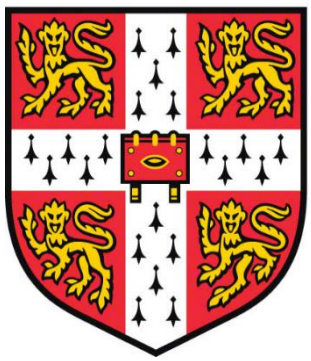
Computer modelling of metabolic adaptions during mitochondrial dysfunction and machine learning to predict novel mitochondrial disease genes



Alexander Gary Smith

Mitochondrial Biology Unit

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Corpus Christi College September 2019

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

It does not exceed the prescribed word limit of 60,000 words as designated by the School of Clinical Medicine Degree Committee.

Computer modelling of metabolic adaptions during mitochondrial dysfunction and machine learning to predict novel mitochondrial disease genes

Alexander Gary Smith

Mitochondria are organelles found in almost every eukaryote and are primarily responsible for generating chemical energy in the form of adenosine triphosphate. This thesis investigates two main causes of mitochondrial dysfunction: mitochondrial toxicity arising from side-effects of drugs; and mitochondrial diseases arising from defects in nuclear-encoded genes.

Novel chemical entities being developed as drug leads are screened for cellular toxicity in which mitochondrial dysfunction is a major cause. However, our lack of understanding of the metabolic adaptations to mitochondrial dysfunction limits the accurate screening of mitochondrial dysfunction for pharmaceutical companies, thus preventing potentially useful drugs from being developed. To further our understanding of these adaptations, I analysed a large-scale metabolomics data set of rats administered a known mitochondrial complex III inhibitor. The analyses revealed many perturbed pathways which can be exploited as biomarkers of mild mitochondrial dysfunction, a condition which is currently clinically undetectable during the drug development process. To direct future studies on mitochondrial dysfunction, a multi-organ model of mitochondrial metabolism was generated and used to simulate inhibition of the mitochondrial respiratory complexes. The simulations of complex III inhibition accurately predicted many of the metabolite behaviours identified in the metabolomics analyses and provided theories for their significance. Simulations of the other complexes' inhibitions identified many unique behaviours which can be used to direct future studies, studies which would greatly improve our understanding of the metabolic adaptations and provide higher confidence biomarkers.

Mitochondrial dysfunction is linked to many late onset diseases such as Parkinson's, and inborn errors of mitochondrial metabolism cause severe neurological and

physiological diseases. Patients with suspected mitochondrial disease have their DNA sequenced and analysed. Diagnosis of mitochondrial disease by sequencing requires knowledge of the mitochondrial proteome, which is currently incomplete. A predicted mitochondrial proteome was generated using a support vector machine trained using the abundance of protein localisation data available in the MitoMiner database. The support vector machine identified 442 novel mitochondrional proteins. The current success rate of diagnosing mitochondrial disease using sequencing is currently limited by our inability to filter and prioritise a patient's DNA variants. Patients which do not have a variant in one of the already known mitochondrial disease genes are usually left with over hundreds of potential disease-causing variants. A probability of being disease-causing for each gene in the mitochondrial proteome was generated using two trained neural networks. The networks were trained on a large amount of different data sources for differentiating mitochondrial disease genes including protein-protein interaction network metrics, gene tissue expression and protein evolution. The predicted probabilities allow for better filtering and prioritisation of a patient's variants for candidate disease-causing genes to be experimentally verified. The predicted mitochondrial proteome and their predicted disease-causing probabilities are currently used in an NGS analysis pipeline at the MRC Mitochondrial Biology Unit for diagnosing mitochondrial disease patient samples.

Acknowledgements

I would like to thank my supervisors Dr. Alan Robinson and Dr. Jon Lyon for giving me this opportunity. A particular thank you to Dr. Alan Robinson for his mentorship and understanding throughout. Thanks to Dr. Anthony Smith for his invaluable assistance on all things modelling, his work on the predicted mitochondrial proteome and many other subjects. Thanks to Dr. Cassandra Smith for allowing me to use her work on mitochondrial protein evolution and for her advice on many other topics which have helped me to pass for a biologist. Thanks to all the people at GSK who made me welcome, particularly Jim Armitage who went out of his way to keep the project rolling. Thanks also to BBSRC, the Medical Research Council and GlaxoSmithKline for funding this work.

A personal thank you to all my family and friends who have supported me along my entire journey. Reaching this point would not have been possible without them.

Summary

Mitochondria are organelles found in almost every eukaryote and are primarily responsible for generating chemical energy in the form of adenosine triphosphate. This thesis investigates two main causes of mitochondrial dysfunction: mitochondrial toxicity arising from side-effects of drugs; and mitochondrial diseases arising from defects in nuclear-encoded genes.

Novel chemical entities being developed as drug leads are screened for cellular toxicity in which mitochondrial dysfunction is a major cause. However, our lack of understanding of the metabolic adaptations to mitochondrial dysfunction limits the accurate screening of mitochondrial dysfunction for pharmaceutical companies, thus preventing potentially useful drugs from being developed. To further our understanding of these adaptations, I analysed a large-scale metabolomics data set of rats administered a known mitochondrial complex III inhibitor. The analyses revealed many perturbed pathways which can be exploited as biomarkers of mild mitochondrial dysfunction, a condition which is currently clinically undetectable during the drug development process. To direct future studies on mitochondrial dysfunction, a multi-organ model of mitochondrial metabolism was generated and used to simulate inhibition of the mitochondrial respiratory complexes. The simulations of complex III inhibition accurately predicted many of the metabolite behaviours identified in the metabolomics analyses and provided theories for their significance. Simulations of the other complexes' inhibitions identified many unique behaviours which can be used to direct future studies, studies which would greatly improve our understanding of the metabolic adaptations and provide higher confidence biomarkers.

Mitochondrial dysfunction is linked to many late onset diseases such as Parkinson's, and inborn errors of mitochondrial metabolism cause severe neurological and physiological diseases. Patients with suspected mitochondrial disease have their DNA sequenced and analysed. Diagnosis of mitochondrial disease by sequencing requires knowledge of the mitochondrial proteome, which is currently incomplete. A predicted mitochondrial proteome was generated using a support vector machine trained using the abundance of protein localisation data available in the MitoMiner

database. The support vector machine identified 442 novel mitochondrional proteins. The current success rate of diagnosing mitochondrial disease using sequencing is currently limited by our inability to filter and prioritise a patient's DNA variants. Patients which do not have a variant in one of the already known mitochondrial disease genes are usually left with over hundreds of potential disease-causing variants. A probability of being disease-causing for each gene in the mitochondrial proteome was generated using two trained neural networks. The networks were trained on a large amount of different data sources for differentiating mitochondrial disease genes including protein-protein interaction network metrics, gene tissue expression and protein evolution. The predicted probabilities allow for better filtering and prioritisation of a patient's variants for candidate disease-causing genes to be experimentally verified. The predicted mitochondrial proteome and their predicted disease-causing probabilities are currently used in an NGS analysis pipeline at the MRC Mitochondrial Biology Unit for diagnosing mitochondrial disease patient samples.

Table of Contents

Chapter 1

Introduction

1.1. The fundamentals of mitochondria

Mitochondria are organelles often referred to as the ‘powerhouse of the cell’, a reference to their primary function of generating cellular energy. They can be found in almost every eukaryote where they make up a large portion of the total cellular volume [1]. Eukaryotes which have no mitochondria are believed to have an ancestor related to the mitochondria [2]. Current evidence suggests that these organelles originated from a symbiotic relationship between free-living bacteria, an α-proteobacteria distantly related to the taxa *Rickettsiales* [3], and an archaeal host-cell [4] which took up these bacteria permanently [5], an event which would have occurred at an early stage of eukaryotic evolutionary history.

Mitochondria are rod-shaped organelles that form into a dynamic interconnected network which is under a constant state of fission and fusion. The variable morphology of mitochondria *in vivo* is critical for mitochondrial biogenesis and aids in the mitigation of metabolic and environmental cellular stresses [6]. Dysfunction of the interconnected network has been associated with many neurodegenerative, cardiovascular, endocrine and neoplastic diseases such as Alzheimer’s disease and cancer [7]. Mitochondria have also been shown to form physical contacts with the Golgi apparatus [8] and endoplasmic reticulum [9].

Figure 1.1 shows the standard double membrane structure of a single mitochondrion. The outer membrane is a permeable membrane which separates the cytosol and inter membrane space. Voltage-dependent anion channels (VDAC) allow

free movement of nearly all solutes under a molecular mass of 5kDa across the outer membrane, making the metabolite content of the intermembrane space virtually indistinguishable from that of the cytosol [10].

In contrast, the inner membrane is completely impermeable and separates the inter membrane space from the matrix, an embedded central compartment. A family of proteins labelled the mitochondrial transporters facilitate the hyper selective transport of metabolites across the inner membrane into the matrix [11]. The surface area of the inner membrane is considerably increased by its folded structure protruding into the matrix. Each of the folds, called a crista, is separated from the inter membrane space by tubular structures, known as crista junctions, which tightly regulate the protein and metabolite levels within the cristae [12]. The shape of the cristae is known to affect mitochondrial dependent cell growth efficiency [13].

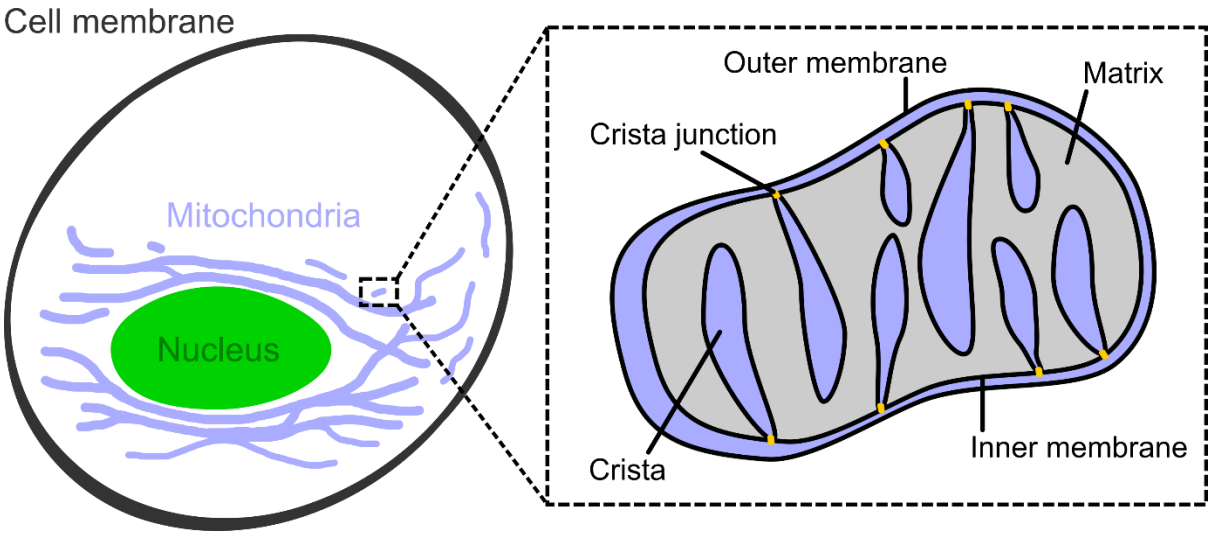


Figure 1.1. Illustration of mitochondria within a cell and the structure of a single mitochondrion.

The production of adenosine triphosphate (ATP), the energy currency of the cell, is an essential function of mitochondria. Lying within the inner membrane is a set of complexes (I-IV) collectively referred to as the electron transport chain (ETC). The ETC, along with ATP synthase, is the site of oxidative phosphorylation, the primary pathway for ATP generation in mitochondria (*Figure 1.2*). ATP synthase produces ATP from adenosine diphosphate (ADP) and inorganic phosphate (P_i) [14]. Its activity is based on the charge ($\Delta\psi$) and proton (ΔpH) gradients, together known as

the proton motive force, across the inner membrane. The controlled re-entry of protons into the mitochondrial matrix by ATP synthase drives rotation of an ATP synthase subunit called the c-ring, which is coupled to the production of ATP from ADP. In animals, approximately 2.7 protons are required to produce one molecule of ATP [15].

Proton motive force is primarily generated by the ETC. Complex I [16], III [17] and IV [18] pump protons out of the matrix across the inner membrane. This process is coupled with the movement of electrons through the ETC from complex I to complex IV. Electrons from NADH (nicotinamide adenine dinucleotide) enter the ETC at complex I (NADH dehydrogenase) and travel down a sequence of iron-sulphur clusters coupled with the pumping of four protons out of the matrix. Hydrophobic ubiquinol within the inner membrane reacts with the electrons to form free moving ubiquinone. Additional ubiquinone is added to the pool via electrons donated by $FADH_2$ (flavin adenine dinucleotide) at complex II (succinate dehydrogenase) [19]. The electrons are then delivered to complex III (cytochrome *c* oxidoreductase) by reduced ubiquinone and four protons per ubiquinone are pumped out of the matrix. The reaction at complex III produces reduced cytochrome *c* which are used to deliver the electrons to complex IV (cytochrome *c* oxidase). Electrons move through complex IV, causing the pumping of four protons out of the matrix, which are ultimately used to reduce oxygen into water using four matrix protons.

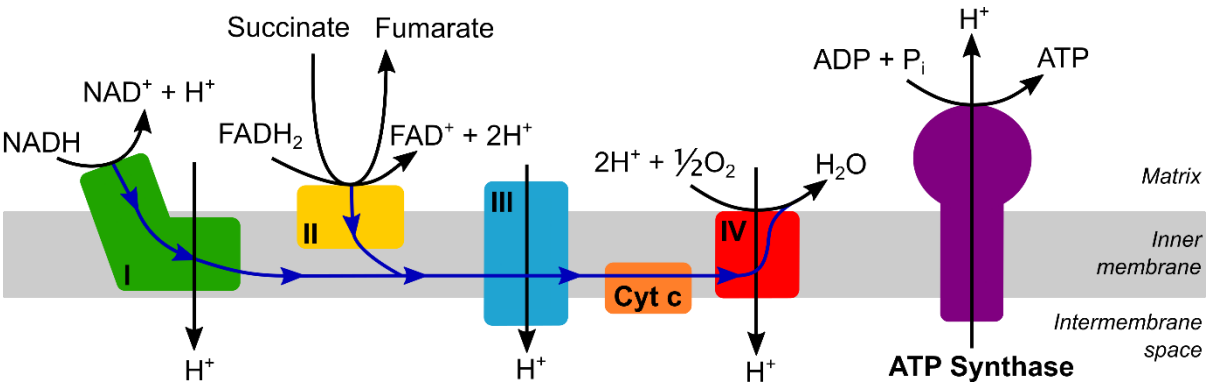


Figure 1.2. Illustration of the electron transport chain (ETC) and ATP synthase reactions involved in oxidative phosphorylation (OXPHOS). The blue arrow indicates the flow of electrons.

The complex II reaction involves the conversion of succinate into fumarate which connects the ETC to the citric acid cycle, a cyclical pathway at the centre of metabolism. The citric acid (TCA) cycle occurs in the mitochondrial matrix and is the end point of a variety of metabolic pathways including glycolysis, fatty acid β-oxidation, ketone body degradation, nucleotide metabolism and various amino acid metabolisms [20]. Mitochondria are also involved in the production of many other essential compounds including steroids [21], haem [22] and iron-sulphur cluster assembly [23].

Other functions of mitochondria include maintaining calcium ion (Ca^{2+}) homeostasis, Ca^{2+} signalling [24] and apoptosis [25]. The mitochondria act as storage sites for cellular Ca^{2+} when cytoplasmic Ca^{2+} concentration is too high. Dramatic and constant increases in mitochondrial Ca^{2+} concentration has been shown to cause the permeabilization of the mitochondrial membranes [24]. The permeabilization causes the release of mitochondrial proteins into the cytoplasm. Several mitochondrial proteins, such as cytochrome *c*, are involved in the mitochondrial apoptotic pathway [26] that, once released into the cytoplasm, induce apoptosis.

1.2. The mitochondrial proteome

The mitochondrial proteome is the library of proteins which at some point in time localise within the mitochondria. Their localisation to the mitochondria implicates these proteins in many essential functions making them an important set of proteins. Inside each mitochondrion is its own circular DNA called mitochondrial DNA (mtDNA), located within its matrix. mtDNA encodes for thirteen subunits of the electron transport chain complexes, two ribosomal RNAs and twenty-two transfer RNAs - a total of thirty-seven genes [27]. Eukaryotes can have multiple different copies of the mitochondrial genome within a single cell, known as mitochondrial heteroplasmy. Current research predicts that the mitochondrial proteome consists of around 1,500 proteins [28], a large portion of which must be nuclear encoded. However, the exact set of mitochondrially localised nuclear encoded proteins is an on-going topic of research, with less than 900 proteins having been experimentally verified and accepted.

Several different experimental methodologies have been used to try and identify proteins which belong to the mitochondrial proteome. MitoMiner is a database which aims to collate all mitochondrial localization data sets for selected species of vertebrates and fungi [29]. The database has a large amount of experimental data sets such as mass spectrometry protein localisation from over 30 different studies and green fluorescent tagging (GFP) data sets from multiple species. Large scale experimental studies of protein localisation such as the Human Protein Atlas (HPA) antibody staining project [30] and the dynamic organellar mapping of HeLa cells [31] are included in MitoMiner. The database also includes computational data sets such as Gene Ontology (GO) annotations [32] and mitochondrial targeting sequence predictions from multiple different programs, such as MitoFates [33]. With the abundance of available protein localisation studies there have been multiple attempts to predict the human mitochondrial proteome. Several older attempts such as MitoProteome [34] and MitoP2 [35] are no longer available. MitoCarta [36], and its successor MitoCarta 2.0 [37], are the most accredited mitochondrial proteomes.

1.3. Mitochondrial disease

The numerous essential functions of the mitochondria mean that even the smallest change in mitochondrial homeostasis can have a drastic effect on humans. Mitochondrial dysfunction is associated with many serious inborn and late onset neurological and physiological diseases. For example, Parkinson’s disease, a generally late onset neurodegenerative disease, has been shown to have a complicated relationship with mitochondrial dysfunction [38]. Research in the 1990’s identified a link between the mitochondria and Parkinson’s when drug-induced complex I inhibition in non-human primates caused a similar phenotype to patients suffering from Parkinson’s disease [39,40]. PINK1 and parkin, two of the major Parkinson’s associated proteins in humans, have been shown to interact in a pathway related to mitochondrial quality control. PINK1, which localises in the mitochondria, recruits cytosolic parkin to the outer membrane of damaged mitochondria to mediate selective autophagy of the damaged organelles [41]. The impairment of either protein leads to the accumulation of dysfunctional mitochondria

causing neurodegeneration in Parkinson's disease [42]. However, the complex relationship between mitochondria and Parkinson's disease is still not fully understood and is currently a hot topic of research in the field of neurodegenerative diseases [43].

Patients with inborn mitochondrial dysfunction usually have many serious and often fatal symptoms. There are many hallmarks of mitochondrial disease, such as apparent maternal inheritance, multiple organs affected, exercise intolerance, and the accumulation of lactic acid in the body, but the diagnosis of mitochondrial disease is often not straightforward or unrecognised. Due to the intricate network of pathways in the mitochondria, the severity of the hallmarks and the clinical phenotype of patients varies from patient to patient depending on the type and location of the individual's mutation(s). For example, Maple syrup urine disease (MSUD) is caused by a mutation to the branched-chain α-ketoacid dehydrogenase complex involved in branched-chain amino acid (BCAA) metabolism [44]. The disease has the clinical hallmark of the accumulation of BCAAs in plasma but the severity of symptoms ranging from a maple syrup odour in urine to Leigh syndrome, a progressive neurodegenerative disorder, is based on which of the complex subunits is mutated [45]. The diagnosis of patients with mtDNA mutations can be particularly difficult as mitochondrial heteroplasmy has been shown to influence clinical phenotype [46].

1.4. Mitochondrial drug safety

Mitochondrial dysfunction can be drug-induced as mitochondria, the centre of metabolism, are highly involved in xenobiotic metabolism. Drug-induced mitochondrial dysfunction causes more subtle, but just as serious, symptoms as mitochondrial disease. Research over the last 50 years has identified many medications that cause mitochondrial dysfunction [47,48]. Mitotoxicity has been attributed to certain anti-diabetics [49], anti-inflammatory [50], anticancer [51] and anti-epileptic [52] drugs. Despite causing mitotoxicity, many medications are still prescribed as treatments as the health benefits of the treatment vastly outweigh the side effects of the generally low levels of mitotoxicity. A famous example being

valproic acid (VPA) which causes hepatotoxicity but remains the world's most prescribed medication for epilepsy and bipolar disorder [53].

Mitochondria are also the intended targets of many drugs. Cancer cells alter their metabolism to promote growth by increasing glucose uptake and promoting glycolysis, the conversion of glucose into ATP, known as the Warburg effect [54]. Potential cancer treatments have been developed which aim to disrupt cellular metabolism by inhibiting the ETC of mitochondria in cancer cells to prevent uncontrolled proliferation and induce apoptosis [55]. As mitochondria are found in almost every eukaryote, drugs have been developed which target the mitochondria of invasive and harmful species to humans. These target a wide range of species such as parasites [56] which aim to prevent the function of the mitochondria and kill off the invasive species.

Regardless of any drug's intended target, pharmaceutical companies must screen every prospective drug for mitotoxicity early in the drug development process. The issue with current methods for screening partial mitotoxicity is that partial mitotoxicity does not have a large effect on the healthy animals used for pre-clinical testing, resulting in the need for a large amount of animal testing. In addition, the drugs may not affect healthy humans during early clinical trials, but severely affect elderly and infirm patients who have compromised mitochondrial function. There is therefore a need to recognise the signs of even weak mitotoxicity during early phase clinical trials and any patient undergoing any form of treatment should be monitored for mitochondrial dysfunction as even the most commonly used drugs, such as Ibuprofen [57], are known to cause low levels of mitotoxicity.

Chapter 2

Metabolic adaptations to mitochondrial dysfunction

2.1. Introduction

2.1.1. The drug development process

The cost of developing a new drug which reaches the market is between 92 and 884 million USD [58]. A major contributor to this huge cost is the high attrition rate at late stages of the drug development process. In a study of 812 compounds put forward as drug candidates by four major pharmaceutical companies, 390 of those reached clinical trials but only two progressed to the final stage, Phase IV [59]. Identification of drug candidates begins by high-throughput screening of thousands of compounds which go through multiple *in vitro* assays and optimisations until only a couple of candidates remain. Lead candidates are then tested preclinically for safety concerns before entering clinical trials [60]. In the study, the reason for the termination of the drug development process was given for 605 drugs where 59% of these were terminated preclinically [59]. Non-clinical toxicology accounted for the largest number of terminations (40%) with 89% of these being at the preclinical stage [59]. An increase of 10% in the efficiency of preclinical screening and termination of lead candidates which later fail in clinical trials would save an estimated 100 million USD in development costs per drug [61].

Non-clinical toxicology testing assesses the potential adverse effects caused by a drug in humans. Lead compounds which are flagged preclinically for toxicity are

investigated thoroughly for mitotoxicity due to the severe symptoms associated with mitochondrial dysfunction. Drugs are initially tested *in vitro* in cells grown in galactose to promote OXPHOS activity [62], and in isolated mitochondria. Mitochondrial function is assessed by various assays such as monitoring oxygen consumption, investigating membrane potential and investigating mtDNA replication or depletion [63]. Candidates which pass safety standards are then tested *in vivo* in healthy rats or mice to assess the potential dose response in humans. The animals are monitored for the clinical signs of mitotoxicity such as changes in core body temperature, irregular breathing patterns and changes in behaviour. The dose response of suspected mitotoxic drugs are assessed *in vivo* using various assays such as measuring arterial blood gas levels of O_2 and CO_2 and measuring lactate/pyruvate ratios of plasma. Risk assessment of the drug is performed based on the *in vitro* and *in vivo* screenings which decides on the advancement or termination of the drug development process.

2.1.2. The problems with current methods

Early identification of mitotoxicity and subsequent risk assessment is a critical step in the drug development process and seeks to help reduce the high attrition rate and costing. However, difficulties in correlating the *in vitro* and *in vivo* results make mitotoxicity screening an expensive and time-consuming process. Drug-induced mitochondrial dysfunction is generally centered around the high energy tissues, such as brain and liver, but drug concentration and activity varies between tissues. Mitochondria are also believed to have an underlying reserve capacity allowing them to adapt to certain stresses *in vivo* [64]. These *in vivo* features are absent *in vitro* causing the discrepancy between *in vitro* and *in vivo* responses, resulting in the need for extensive *in vivo* studies that are both expensive and a potentially avoidable use of animal models. Current *in vivo* assays are performed on isolated liver mitochondria taken from the treated animals, but the process of isolating the mitochondria from the liver can wash away the effect of the drug. These issues, compounded by the general lack of understanding of mitochondrial adaptations under specific stresses, make the current methods for mitotoxicity screening both

inefficient and inaccurate at identifying the precise cause of any drug-induced mitochondrial dysfunction.

Mitochondrial reserve capacity is the amount of additional ATP that can be produced by OXPHOS when the energy demands of a cell are suddenly increased. Cells in a natural state of high energy demand, such as those under metabolic stress, will have a reduced mitochondrial reserve capacity. The age-related decline in OXPHOS efficiency has been implied to cause an age-related decline in mitochondrial reserve capacity [65]. Therefore, the healthy, young rats or mice used for the *in vivo* testing have a relatively high mitochondrial reserve capacity. This causes a steep dose-response curve where only a minor difference in the concentration of an administered drug can be the difference between no clinical signs of mitotoxicity and animal fatality; the clinical signs of mild mitotoxicity are non-existent. However, the intended human recipients of the drug will likely have a reduced mitochondrial reserve capacity due to age and cellular stress from illness. The mild mitotoxicity which is largely undetectable in the animal models may cause clinical effects on the intended patients.

The first step towards solving both these issues is furthering our understanding of the metabolic adaptations which occur during mitochondrial dysfunction, both in general and in relation to specific types of mitochondrial dysfunction.

2.1.3. Biomarkers of mitochondrial dysfunction

A biomarker (“biological marker”) is a molecule found in tissue or bodily fluids which indicates either a normal or abnormal process of a disease or condition [66]. They are generally measured in readily available bodily fluids, such as blood or urine, to detect or monitor sub-clinical disease, and to monitor the response to treatments. For example, prostate cancer can be detected and monitored by the level of prostate-specific-antigen (PSA) found in blood plasma [67]. The identification of a plasma biomarker of mitochondrial dysfunction would enable the identification of mitotoxicity at an early stage of drug development using a relatively non-invasive technique, which is both inexpensive and would not require a large number of animal

models. The biomarker could be used to monitor the level of mitotoxicity to investigate the dose-response curve without the need for clinical phenotypes and enable better risk assessment of a drug's potential use in humans. In addition, any patient undergoing treatment with a drug believed to cause mitotoxicity can be monitored to ensure that damaging levels of mitochondrial dysfunction are avoided. This is especially important for patients taking a concoction of medications with an already reduced mitochondrial reserve capacity.

Advances in technology and experimental procedures over the last decade has enabled the rise of 'omic' based studies; genomics, transcriptomics, proteomics and metabolomics [68]. Metabolomics focuses on the measurement of metabolites, the small molecules involved in metabolism, most commonly performed by either mass spectrometry [69] or nuclear magnetic resonance (NMR) spectroscopy [70]. The measurement of every metabolite, known as the metabolome, generates a global metabolic profile of a cell, tissue or organism under a certain set of conditions. The metabolome is sensitive to changes in the transcriptome and proteome as it is the final downstream product of gene transcription and is the closest measurable level to biological phenotype. As such, metabolomics has emerged as a leading method of biomarker identification [71] for many diseases and conditions such as cancer [72].

Metabolomics only gives a snapshot of a system's metabolic activity as metabolic profiles are highly dynamic. The human metabolome is expected to be more than 19,000 metabolites [73]. For biomarker identification, the size of the metabolome and the dynamic nature of metabolic profiles necessitates a large set of metabolites to be measured to accurately infer metabolic activity. The approach of measuring the largest possible subset of the metabolome in an unbiased way to investigate metabolic activity is often referred to as untargeted metabolomics [74]. Measuring the metabolic profile at different time points is also essential to unravelling the dynamic nature of metabolism. Once the metabolic activity of a system under certain conditions is understood, metabolites which are intricately involved in the metabolic activity or adaptions can be identified and used as biomarkers.

Recent attempts to identify metabolic biomarkers of mitochondrial dysfunction was performed using small scale untargeted metabolomics on patients with a monogenic form of Leigh syndrome [75]. Leigh syndrome is a severe neurological disease

caused by a loss of function mutation in *LRPPRC* which results in a decline in mitochondrial respiration and has many shared hallmarks with mitochondrial disease. The study measured a relatively small subset of metabolites (*n* = 143) in a cohort of patients and compared them to age matched controls. The study identified many significant metabolites using simple statistical methods. The most notable biomarkers included increased plasma levels of β-hydroxybutyrate, a ketone body used as an alternative fuel source, lactate, the most common hallmark of mitochondrial disease, pyruvate, an intermediate in glycolysis, and various fatty acid carnitines, the molecular form of fatty acids when they are being transported into cells for metabolism. The small subset of measured metabolites and the relatively unknown mechanism causing the mitochondrial dysfunction prevents the study from being utilized in trying to understand the precise metabolic adaptations which occur during mitochondrial disease. Furthering our understanding of the metabolic adaptations which occur during mitochondrial dysfunction by performing large scale untargeted metabolomics in other disease models will help filter out the list of significant metabolites identified in this study and potentially lead to biomarkers for mitochondrial disease which has been experimentally verified in humans.

2.1.4. Analysis of metabolomics data sets

Untargeted metabolomics produces large data sets which are usually compared with a control group. Statistical analysis of the comparison identifies the metabolites that are significantly different between the two groups. The large data sets are pre-processed before statistical analysis to mitigate the large difference in concentration level between different metabolites which aims to improve the biological interpretability of the data [76]. The most common methods for identifying significantly different metabolites are the dimensionality reducing techniques of principle component analysis (PCA) and partial least squared discriminant analysis (PLS-DA) [77].

PLS-DA is the categorial adaptation of regression modelling that has become the primary method for metabolomics analysis in recent years. For example, PLS-DA was used to analyse metabolomic data collected from patients suffering with

Parkinson's disease [78] and esophageal cancer [79] with the aim of identifying biomarkers. It is a supervised machine learning technique that generates a model that can classify unknown samples. The classification power of PLS-DA models is no better than simpler methods, but PLS-DA comes with the added ability to quantitively identify the important variables [80]. A variance importance in projection (VIP) score is assigned to each variable which represents the variable's classification power. The assignment of VIP scores makes PLS-DA a useful tool in exploratory analysis and, in the case of metabolomics, enables the identification of significant metabolites. However, metabolomics data sets generally have a large fold difference between the number of samples (*n*) and the number of variables (metabolites). This has been identified as an issue for PLS-DA leading to high error rates and the unreliable classification of samples [81]. Therefore, PLS-DA should be used in conjunction with other methods when performing exploratory analysis on metabolomics data sets [80].

Support vector machines (SVMs) are a supervised machine learning algorithm [82] which do not suffer from the same drawbacks as PLS-DA. A comparative study on the predictive power of PLS-DA versus SVMs identified SVMs as the superior choice for metabolomics studies [83]. Exploratory analysis can be performed using recursive feature elimination (RFE) with SVMs. Recursive feature elimination (RFE) is used when generating SVM models to determine the minimal set of features, i.e. metabolites, of the data set that generate the most accurate SVM model based on sample classification. The algorithm starts by generating a classification model for each individual metabolite. The metabolite that generates the best model, based on classification accuracy, is kept and its accuracy is stored. The algorithm then generates a new model using the best performing metabolite combined with one additional metabolite, for every pairwise combination of metabolites. The two best performing metabolites are kept, and their accuracy is stored. The algorithm continues this recursive process of adding a single metabolite to the previous set of best performing metabolites until all metabolites are added. Assessing how classification accuracy changes as metabolites are added identifies the smallest subset of metabolites that generate the most accurate classification model, along with a ranking of the metabolites based on their classification power. In data sets where sample numbers are small, and variables have high variance, SVM-RFE had

great success in identifying the important variables for classification, making it the superior tool for biomarker identification in metabolomics studies [84]. Recent studies have used SVMs to explore the metabolomics of patients with ovarian cancer [85] and celiac disease [86].

Both PLS-DA and SVMs can be used in conjuncture with cross-validation which helps to avoid model overfitting, where a model is too tightly tuned to a data set causing the model to predict poorly on new unclassified data. Cross-validation involves partitioning the data into two sets – a training set and a validation set – and training the machine learning algorithm with the training set whilst using the validation set to assess model performance. The process of partitioning and creating/assessing a model is repeated multiple times depending on which type of cross-validation method is used. The average accuracy of each partition is then used to assess the overall model performance, known as cross-validation accuracy.

The combination of PLS-DA, SVM-RFE and other simpler statistical methods (such as hypothesis testing and PCA) should be used when performing exploratory analysis on metabolomics data. By combining the results of each method, the most robust subset of significant metabolites can be identified and examined for potential biomarkers.

2.1.5. Chapter summary

In this chapter, I describe the data pre-processing and statistical analysis performed on a metabolomics data set collected from a study on drug-induced mitochondrial complex III inhibition in rats. The identified metabolic adaptations are discussed in the context of mitochondrial dysfunction, and potential biomarkers for both mild and complete mitochondrial dysfunction are identified and discussed.

2.2. Methods

2.2.1. Metabolomics study outline

GSK932121A is a potent inhibitor of mitochondrial complex III in *Plasmodium falciparum* and began development as an antimalarial drug. However, the drug showed an affinity for human mitochondrial complex III and thus mitotoxicity [87], causing its development to be halted. Subsequently, GSK932121A was used in a large-scale metabolomics study focused on investigating mitochondrial dysfunction *in vivo*. The study used twenty female rats separated into three groups: a control group of six rats; a 'low dose' group of seven rats; and a 'high dose' group of seven rats. The low dose group were administered 12.5 mg/kg, known to be non-lethal but cause mitotoxicity, and the high dose group were administered 50 mg/kg, known to be lethal if the animals were left for greater than four hours, both via intraperitoneal injection to the body cavity. Pre-dose plasma samples were taken from all rats, followed by plasma samples collected at 0.5, 1, 2 and 4 hours after dosage. Liver samples were collected as a terminal procedure at 4 hours, giving a total of 6 samples collected for each rat. All experiments and sample collections were performed by GlaxoSmithKline, and Metabolon, Inc. generated the metabolomics data for each sample in the form of peak intensities using four different techniques: UPLC-MS/MS with negative ion mode electrospray ionization, UPLC-MS/MS with positive ion mode electrospray ionization, LC polar platform and GC-MS. All animal studies were ethically reviewed and carried out in accordance with Animals (Scientific Procedures) Act 1986 and the GSK Policy on the Care, Welfare and Treatment of Animals.

2.2.2. Data set quality control

Administration of the drug by intraperitoneal injection to the body cavity can result in 'gut dosing' which causes the drug to remain within the gut of the animal and greatly reduces the exposure of the animal's high energy organs to the drug. The terminal liver samples were all tested for drug concentration where samples with abnormally low drug concentration were removed from the study.

Metabolites with greater than 35% of measurements missing (over case and control) were removed from the data set as inference would be weak over such a small sample set. Missing measurements for the remaining metabolites were imputed from the known measurements using the nearest neighbour algorithm, performed using MATLAB [88]. Imputation of missing values was necessary as support vector classification models cannot deal with empty values. Z-scores of every metabolite in each sample were generated using standardisation over the complete data set (control, low dose and high dose samples) to evaluate class separation and identify potential outliers, performed using R [89].

2.2.3. Statistical methods

Principle component analysis was performed after normalization (mean of zero and standard deviation of one) of the data sets to perform initial investigation of sample separation using R. The large metabolomics data set was then separated into a control and low dose data set and a control and high dose data set for subsequent analyses. Initial exploration of the two data sets was done by hypothesis testing to begin identifying which metabolites were significantly different between case and control. Welch's two-sample t-test [90] was used to calculate p-values that were then corrected for multiple testing using the Benjamini-Hochberg false discovery rate [91] (q-values). Both steps were performed on log-transformed data by utilizing the Python [92] 'scipy' [93] and 'statsmodels' [94] packages. All correlation tests performed in the study were done using a two-tailed Spearman's rank-order correlation test [95] and performed in R.

2.2.4. Machine learning classification models

For the control and low dose data sets, SVMs were used to generate a support vector classification (SVC) model, which predicted the class membership of each data sample. In this study, a linear kernel was used with RFE and cross-validation to generate SVC models using a c-parameter of 0.01 and squared hinge loss [96]. As

SVMs were used in combination with other analytical methods, square hinge loss was used to avoid attaining a sparse solution. Various values of c-parameter were explored, only the extremities made a large difference on the accuracies of the model with 0.01 providing the best accuracy with a reasonably small vector margin. L2 regularisation was used to address the collinearity problem of metabolites and reduce overfitting. The SVC models were generated using stratified 6-fold cross validation, meaning each cross-validation step withheld one control sample and one or two low dose samples. Seven SVC models were trained for each of the 6 time points to ensure that every possible combination of control and low dose sample was used in the validation set at least once during cross-validation. The models were generated on data standardised to a mean of zero and standard deviation of one. The methods were implemented in Python using the 'sklearn' package [97].

A single PLS-DA model was generated for each of the 6 time points in the control and low dose data set. PLS-DA models and VIP scores were generated using MATLAB v8.6.0 and PLS-Toolbox v8.1 [98]. All data were standardised to a mean of zero and standard deviation of one and leave-one-out cross-validation, where one sample is withheld at each cross-validation set, was employed to avoid overfitting.

Evaluation method	**Low significance: 1 point**	**Medium significance: 2 points**	**High significance: 3 points**
Hypothesis testing: q-value	$0.3 \geq x > 0.15$	$0.15 \geq x > 0.10$	$x \leq 0.10$
SVM RFECV: Mean rank	$200 \geq x > 100$	$100 \geq x > 25$	$x \leq 25$
PLS-DA: VIP score	$2.0 \leq x < 2.5$	$2.5 \leq x < 3.5$	$x \geq 3.5$

Table 2.1. The criteria for consistency points for each of the three analyses performed in this study based on different significance levels.

2.2.5. Metabolomics network visualisation

The network representation of central metabolism was generated in Cytoscape [99] with metabolites as the nodes and reactions as the edges. All information was taken from KEGG [100] and utilized the KEGG reaction and compound IDs. The colour of a node represents the fold-change of the metabolite between case and control, red being increased (higher in case than control) and blue decreased (lower in case than control). Nodes coloured in white were measured but found not significantly different, whilst grey nodes were not measured at all. The size of the node was proportional to the metabolites consistency level between control and low dose based on a cumulative scoring system evaluated over the three analyses performed in this study (*Table 2.1*). A higher score meant greater consistency and a larger node size. Metabolites were placed into consistency levels ranging from 1 to 9, with 9 being the highest consistency which indicated that all three analyses identified the metabolite as highly significant.

2.3. Results

2.3.1. Clustering of samples

The mitochondrial drug exposure level of all the rats administered the lower dosage of 12mg/kg was relatively uniform (*Figure 2.1a*) aside from two samples which had abnormally low concentrations of the drug and thus removed from further analysis. Two of the high dose samples administered 50mg/kg, sample 18 and 20, had exposure levels like that of the low dose sample and clustered with the low dose samples based on hierarchical clustering (*Figure 2.1b*). These two samples were therefore moved into the low dose group leaving six control rats, seven low dose rats and only four high dose rats. The rest of the high dose samples showed a larger variation in exposure levels, with sample 17 having a particularly extreme exposure level.

Before analysis the liver and plasma data sets were put through quality controls and imputation, resulting in the liver subset consisting of 844 measured metabolites and the plasma subset consisting of 787 measured metabolites. The extremity of sample

17 was reflective in the plasma 4-hour metabolite Z-scores where it showed the greatest variance among the samples, although this was not as obvious for the liver 4-hour metabolites (*Figure 2.2*). All other samples, including the controls, had a larger variance in liver than plasma at 4 hours, with the low dose group having the lowest variance overall. None of the samples appeared to be outliers so all remaining samples were used for further analysis.

For the plasma predose, 0.5-hour and 1-hour samples, none of the three groups showed any meaningful separation on a PCA plot (*Figure 2.3*). At two hours, the control group clustered away from the two dosage groups, which both clustered together showing no clear separation. The plasma 4-hour samples had clear separation between each of the three groups along with relatively small confidence ellipses. The liver 4-hour samples had a modest amount of separation where the high dose samples were the most differentiable from the rest, but the low dose and control samples clustered near to each other.

2.3.2. Identifying statistically significant biomarkers by hypothesis testing

The full results of all analyses, along with each metabolite's assigned consistency level can be found in *Supplementary File 2.1*. In both the control vs. low dose and control vs. high dose data set, hypothesis testing for the plasma predose, 0.5-hour and 1-hour time points identified almost no metabolites as significantly different after adjusting for multiple testing, using a very lenient significance threshold of q = 0.3 (*Figure 2.4, Table 2.2*). In contrast, the plasma 2-hour and 4-hour time points had many metabolites found to be significantly different in both data sets. For the control vs low dose data set, plasma 2-hours had 202 metabolites with a q-value less than or equal to 0.3 while plasma 4-hours had 192 metabolites. The control vs high dose data set identified 174 metabolites at plasma 2-hours and 351 metabolites at plasma 4-hours. In the liver 4-hour samples, the control vs low dose data set only identified 4 metabolites as significant whilst the control vs high dose data set had 192 metabolites significantly different between case and control. A lenient value for significance was selected as a larger subset of metabolites was needed to understand the metabolic adaptations occurring due to mitochondrial dysfunction, allowing for the identification of a more biologically relevant biomarker. The higher

false discovery rate was counteracted by using the results of the hypothesis testing in combination with the other statistical methods.

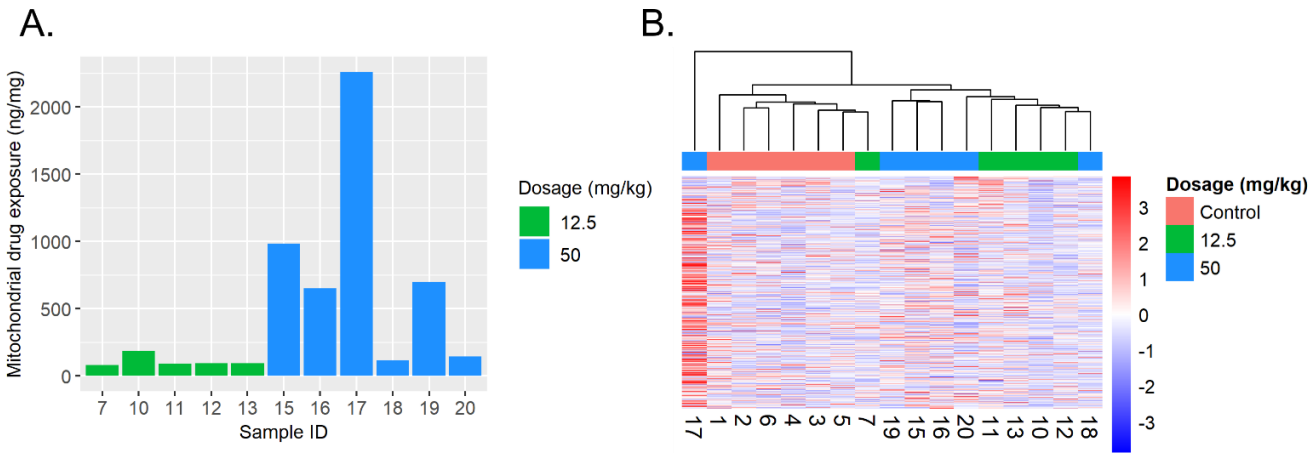


Figure 2.1. A) Bar chart to show the measured drug exposure level in each sample. **B)** Hierarchical clustering of the samples at the plasma 4 hour time point.

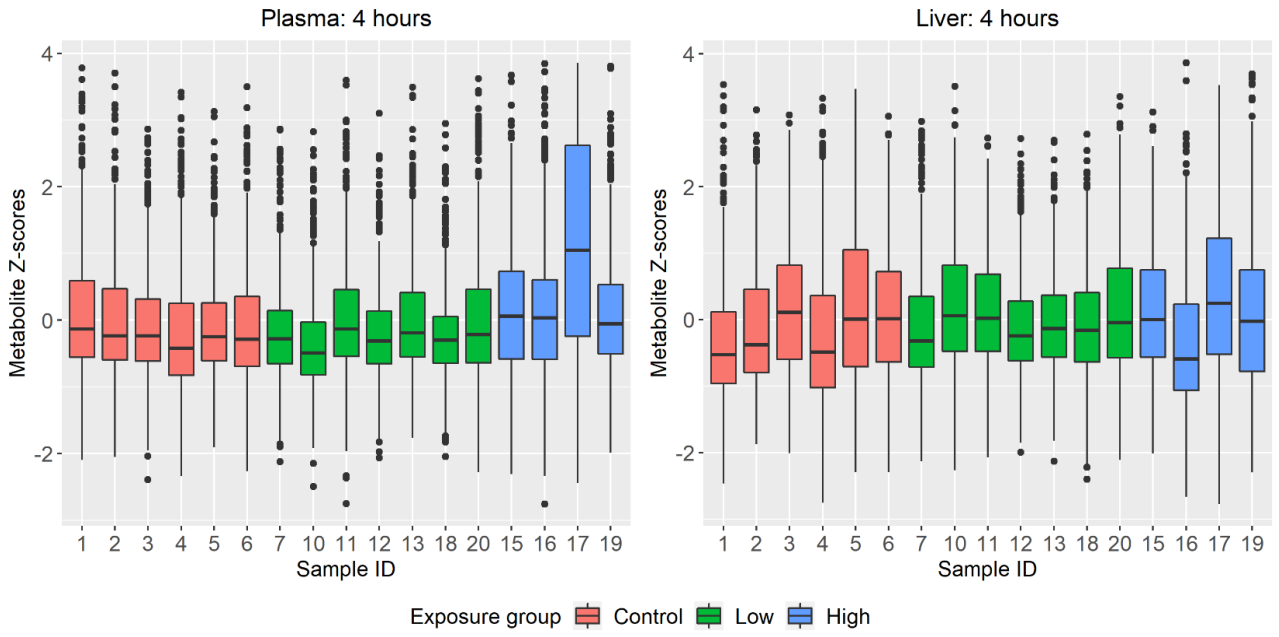


Figure 2.2. Box-plots of metabolite z-scores for each sample.

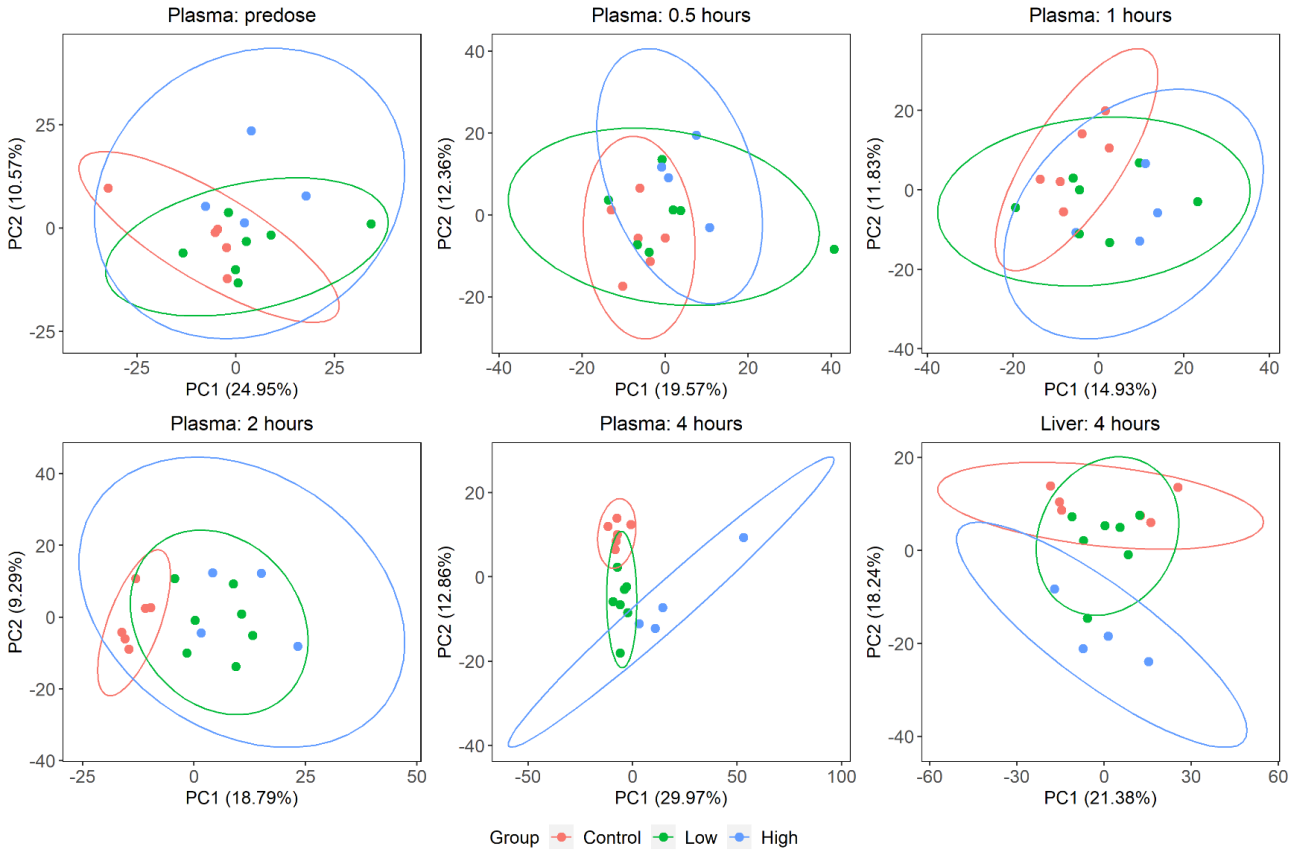


Figure 2.3. Principle component analysis (PCA) plots for each of the time points. Each ellipsoid represents the group's 95% confidence interval.

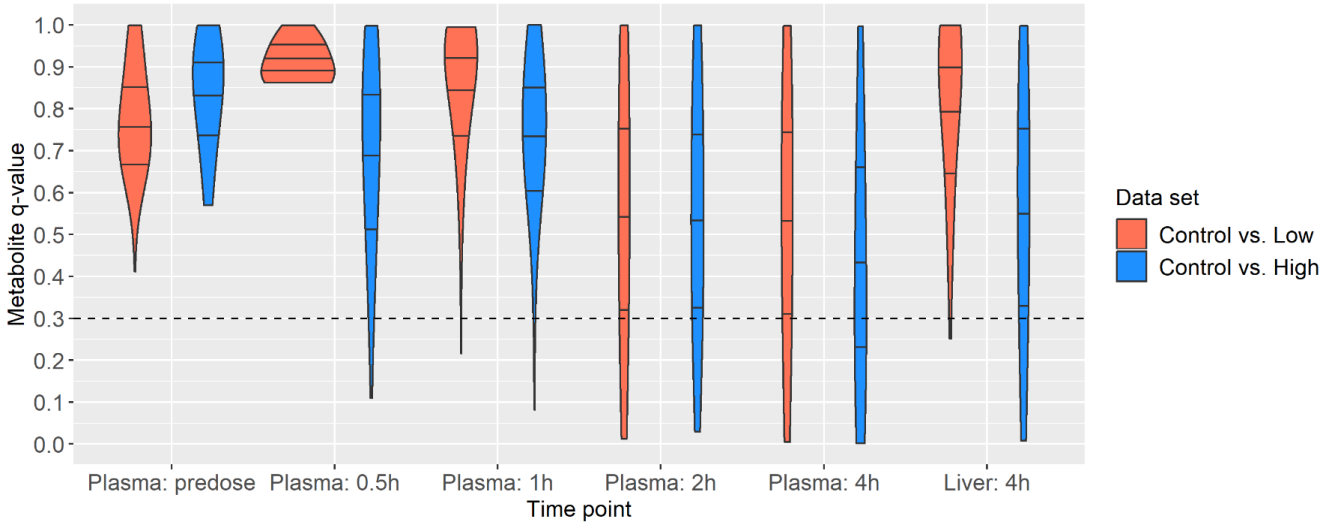


Figure 2.4. Violin plots of the q-values for all metabolites at each time point for both data sets, control vs low dose and control vs high dose. The lines on each violin indicate the upper quartile, mean and lower quartile. The dotted line indicates the threshold for significance used, q = 0.3.

Dosage time point	Low dose	High dose
Plasma pre-dose	0	0
Plasma 0.5 hour	0	8
Plasma 1 hour	3	3
Plasma 2 hour	202	174
Plasma 4 hour	192	351
Liver 4 hour	4	192

Table 2.2. The number of metabolites for each time point which were identified as significantly different between case and control based on hypothesis testing ($q \leq 0.3$).

2.3.3. Identifying statistically significant biomarkers by support vector machine (SVM) recursive feature elimination with cross validation (RFECV)

Seven SVM models for classification were generated at each time point in the control vs. low dose data set. A generalized additive model, a type of generalized linear model, was fit to each of the time points using the cross-validation accuracy of the seven models at every RFECV increment, shown in *Figure 2.5*. The cross-validation accuracies of the plasma predose, 0.5-hour and 1-hour time points remained below 75% accuracy throughout the RFECV process. In combination with the PCA and hypothesis testing results, these three time points were found to be not significantly different between case and control and were discarded from further analysis.

The remaining three time points consistently achieved greater than 75% cross-validation accuracy, with the plasma 2-hour and 4-hour models achieving over 85% accuracy. The RFECV process ranked each metabolite based on its classification power, a lower rank indicating higher classification power. These values were averaged over the seven SVM models per time point resulting in 107 metabolites

having a mean rank less than or equal to 100 at plasma 2-hours, 106 for plasma 4-hours and 134 for liver 4-hours. Machine learning models could not be generated for the control and high dose data set as the number of high dose samples (*n* = 4) was too small to ensure the reliability of the models.

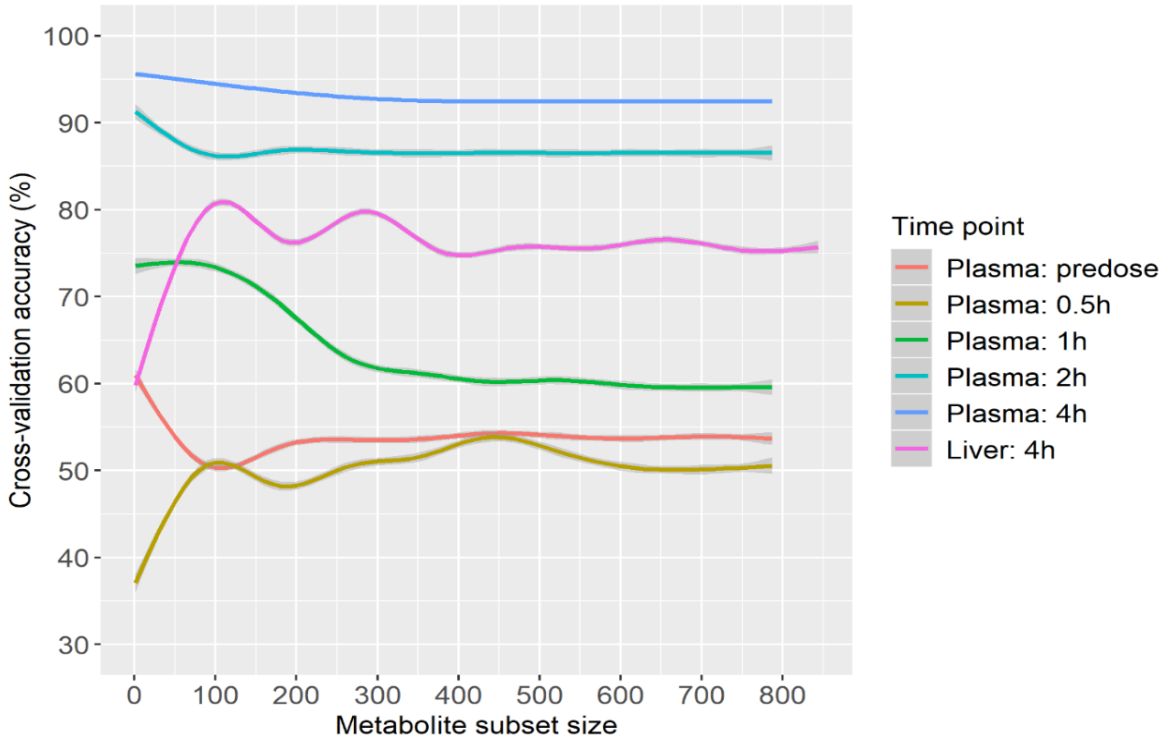


Figure 2.5. Cross-validation accuracies of the support vector classification models generated for each time point during the recursive feature elimination process. Each line represents a generalized additive model fit using the cross-validation accuracies of the seven SVC generated for each time point.

2.3.4. Identifying statistically significant biomarkers by partial least squares discriminant analysis (PLS-DA)

PLS-DA models were generated for the plasma 2-hour, plasma 4-hour and liver 4-hour time points for the control vs. low dose data set. All three of the models were generated using two latent variables and resulted in clear separation of the two groups (*Figure 2.6a*). Both the plasma 2-hour and plasma 4-hour models had a Root Mean Squared Error Cross Validation (RMSECV) of less than 0.3 along with an R^2 greater than 0.67, indicating that both were good models for classification with minimal overfitting [101]. The liver 4-hour model had an RMSECV of 0.41 along with an R^2 of 0.33 which does not indicate a particularly good model for classification.

However, this coincides with the results of both the SVM and hypothesis testing indicating that the liver 4-hour time point was only marginally significantly different between case and control. In spite of this result, the model was kept for further analysis.

VIP scores for every metabolite were taken from each of the models, a metabolite with a VIP score greater than 1 had a significant amount of classification power, with a higher score meaning higher classification power (*Figure 2.6b*). The plasma 2-hour model identified 299 metabolites as significant (VIP ≥ 1), with 130 of these being highly significant (VIP ≥ 2). In the plasma 4-hour model, 306 metabolites were significant with 122 of these being highly significant, whilst the liver 4-hour model had 301 metabolites significant with 132 of these being highly significant.

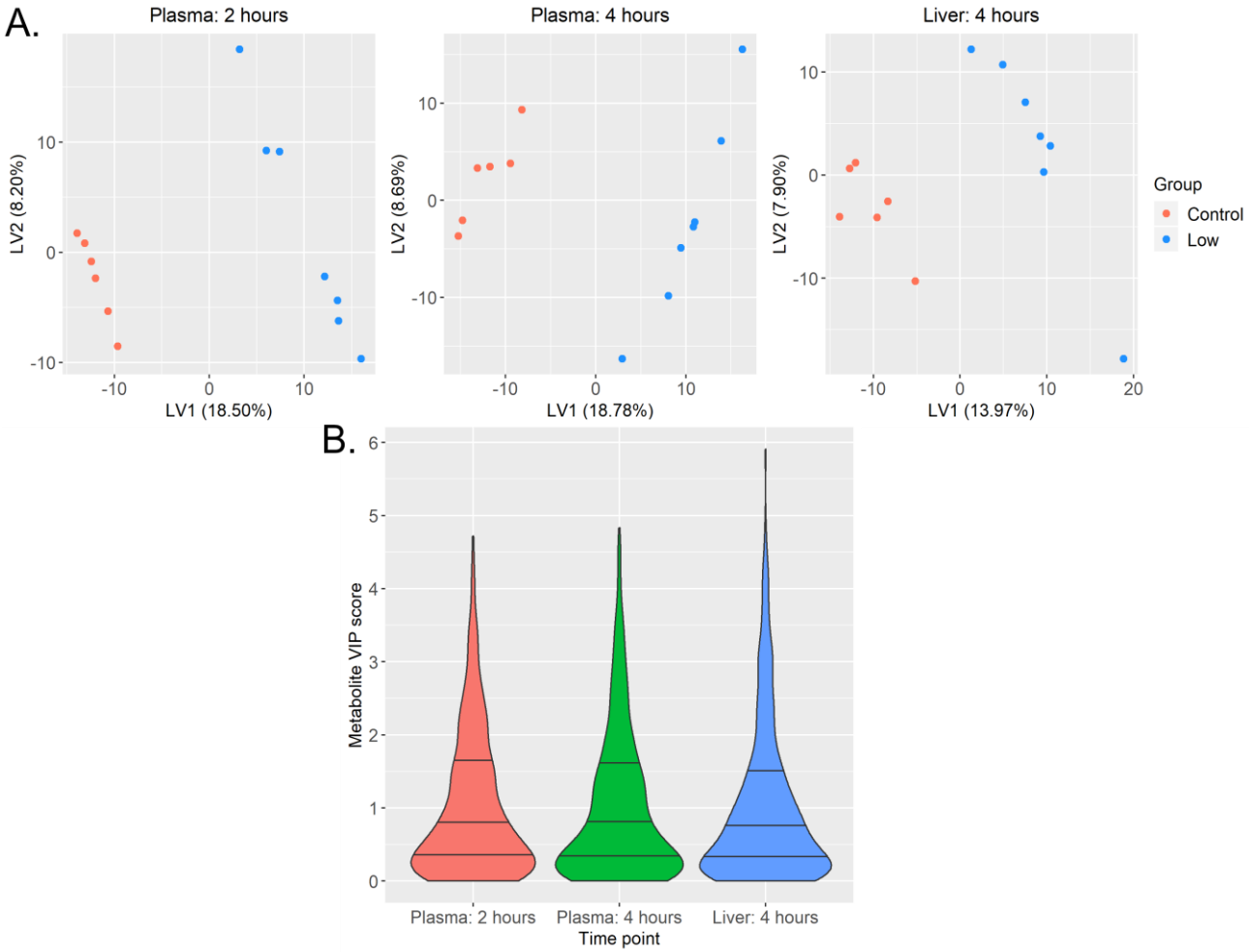


Figure 2.6. A) Latent variable (LV) plots of the PLS-DA models generated for the plasma 2-hour, plasma 4-hour and liver 4-hour time points. **B)** Violin plots of the VIP scores for each of the PLS-DA models. The lines on each violin indicate the upper quartile, mean and lower quartile.

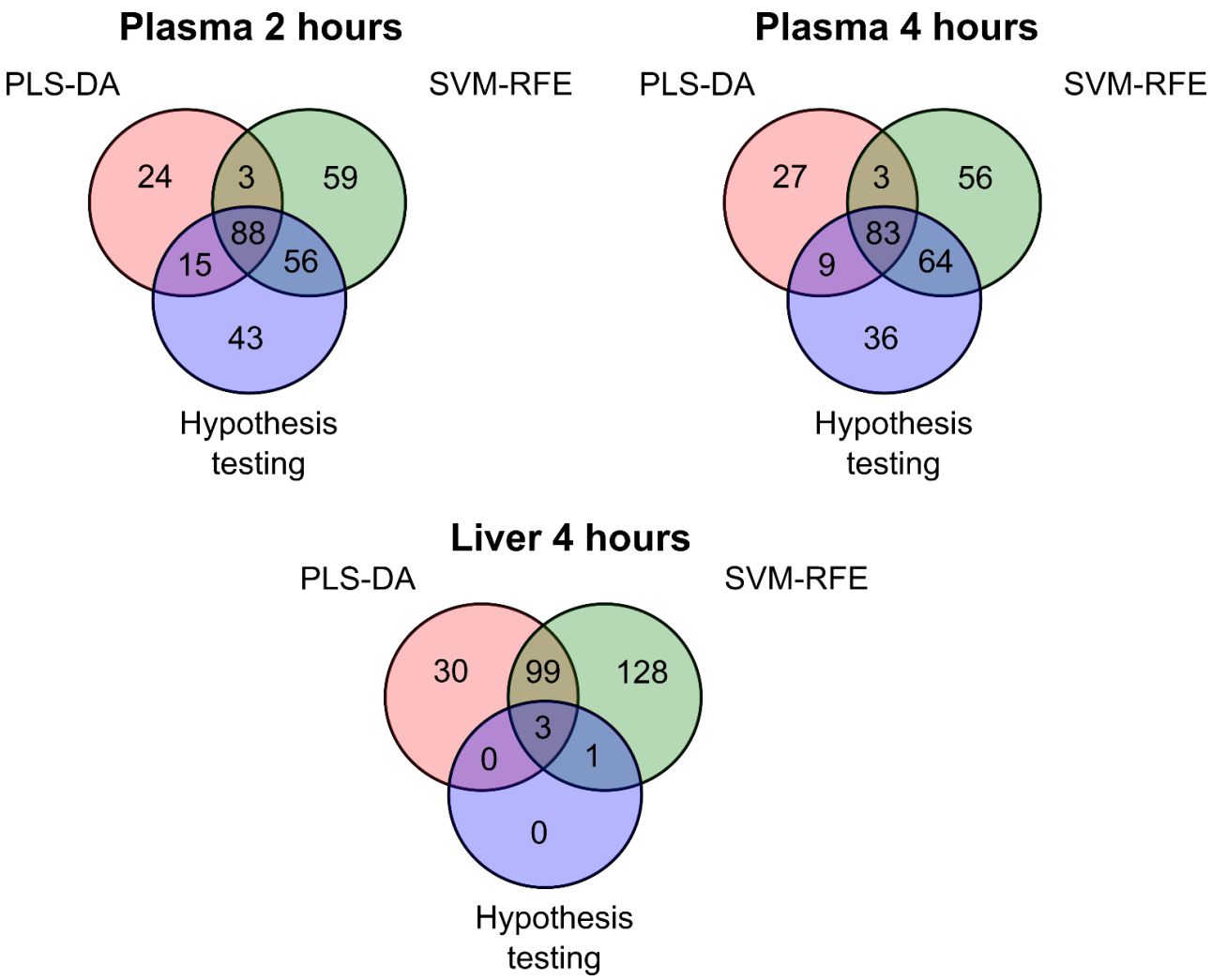


Figure 2.7. Venn diagrams showing the number of significant metabolites identified by each method for the low dose rats. PLS-DA with VIP ≥ 2.0, SVM-RFE with mean rank ≤ 200 and hypothesis testing with q-value ≤ 0.3.

2.3.5. Identifying a set of biomarkers for mild mitochondrial dysfunction

Using a VIP threshold of 1 at each time point identified a subset of metabolites found significantly different between the control and low dose samples which, along with their fold change, could have been used to interpret the metabolic adaptations occurring due to mild mitochondrial dysfunction. However, due to the problems with PLS-DA, this choice of threshold was assessed by comparing the results of the PLS-DA modelling with the hypothesis testing and RFECV results. The cross comparison between the number of metabolites found significant for each of the methods is shown in *Figure 2.7*.

Figure 2.8a shows each metabolite's VIP score plotted against its q-value from hypothesis testing, along with a fitted generalized additive model. Spearman's rank-order correlation identified significant correlation between these two results for all the time points (plasma 2-hour: Rho = -0.95, plasma 4-hour: Rho = -0.95, liver 4-hour: Rho = -0.83). Based on the generalized additive models, a VIP score of 1 corresponded to a q-value of around 0.5 in the plasma time points and 0.85 in the liver time point, which were extremely high. If the subset of metabolites was selected based on these q-values, the subset would be riddled with false positive results. The q-values reduced quickly as the VIP scores increased, with a VIP score of 2 being much closer to the 0.3 q-value significance threshold used earlier in the study. The higher q-values present in the liver 4-hour time point were accepted under the assumption that this time point was only minorly significantly different between control and low dose.

Figure 2.8b shows similar analysis for metabolite VIP scores and metabolite mean RFECV rank, along with its fold-change difference in low dose compared to control. Significant correlation was found between metabolite VIP score and mean RFECV rank for all three time points (plasma 2-hour: Rho = -0.84, plasma 4-hour: Rho = -0.85, liver 4-hour: Rho = -0.81). The plot showed that, compared to metabolites with a VIP score of 1, metabolites with a VIP score of 2 had on average a much lower mean rank and larger fold-change difference.

Both these results indicated that a VIP threshold of 1 would not be strict enough and would result in a potentially large number of false positives. This in turn would make the interpretation of the metabolite adaptations caused by mild mitochondrial dysfunction particularly difficult, leading to the identification of incorrect biomarkers. The threshold for metabolite significance was thus set at a VIP score of 2. This generated a subset of 130 metabolites for plasma 2-hours, 122 metabolites for plasma 4-hours and 132 metabolites for liver 4-hours. A summary of the statistics for each of these significant metabolites can be found in *Appendix I: Table 1-3*.

The consistency of these metabolites was evaluated based on each of the three analyses performed and ranked using a cumulative points system (*Table 2.1*). The plasma 2-hour time point had a total of 63 metabolites placed into level 5 or greater whilst the plasma 4-hour time point had 60 metabolites (*Figure 2.9*). The liver 4-hour

time point had 41 metabolites in level 5 or greater, although none of these were greater than level 7.

The significant metabolite subsets for each time point were used to generate network heatmaps (*Figure 2.10, Supplementary File 2.2*). The heatmap identified multiple pathways which had clusters of significant metabolites; purine metabolism, pyrimidine metabolism, glutathione metabolism, ketogenesis and glucose/carbohydrate metabolism.

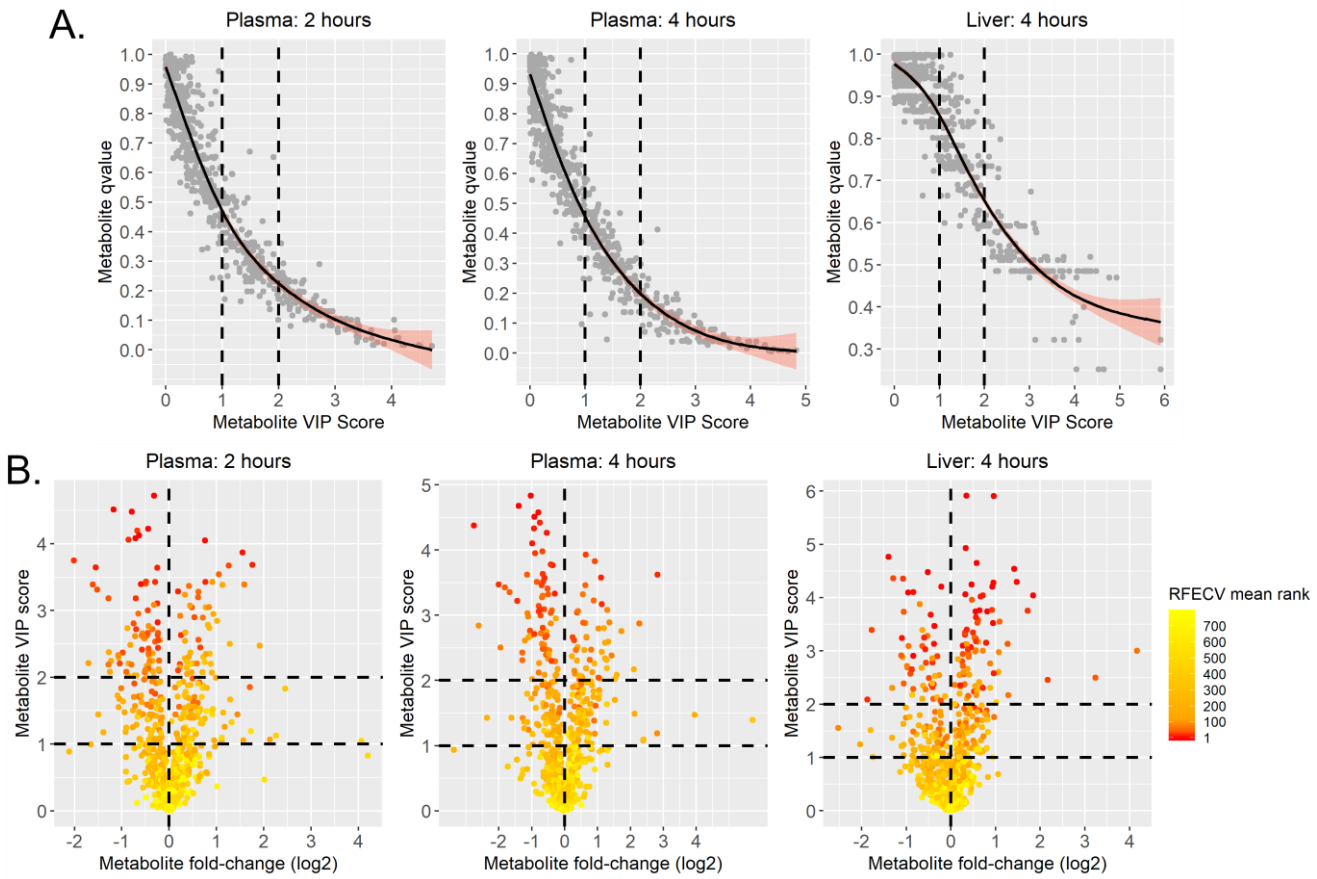


Figure 2.8. A) Scatter plot of a metabolite's q-value against its VIP score with a fitted generalized additive model indicated by the line, along with its confidence interval. **B)** Volcano plots of a metabolite's VIP score against its fold-change difference in low dose compared to control. Each metabolite is coloured by its RFECV mean rank.

2.3.6. Identifying a set of biomarkers for complete mitochondrial dysfunction

The analysis of the control vs low dose data set identified a set of potential biomarkers for mild mitochondrial dysfunction. A biomarker for mild mitochondrial dysfunction can only be a biomarker for complete mitochondrial dysfunction if it is also a biomarker for fatal levels of mitochondrial dysfunction, such as the high dose samples used in this study. In addition, the level of the metabolite must be correlated with the exposure level of the drug to enable the identification of the level of mitochondrial dysfunction occurring. For each metabolite belonging to each of the significant metabolite subsets, the correlation was calculated between the metabolite's measured level and the drug exposure level of the sample that it was collected from, across both the low and high dose samples.

Figure 2.11a shows each metabolite's calculated Rho value, along with its respective control vs high dose q-value. Metabolites with a q-value less than 0.15 and an absolute Rho value greater than 0.5 were highlighted in red, which were the potential biomarkers for complete mitochondrial dysfunction. The plasma 2-hour time point only had two metabolites which fit this criterion, with only one of those having a relatively high consistency level in the control vs low dose data set (*Figure 2.11b*). For plasma 4-hours, there were 20 metabolites which fit the criteria, with five of these belonging to the highest consistency level in control vs low dose. Finally, the liver 4-hour time point had 32 metabolites which fit the criteria and 10 of these were in a consistency level of 5 or greater.

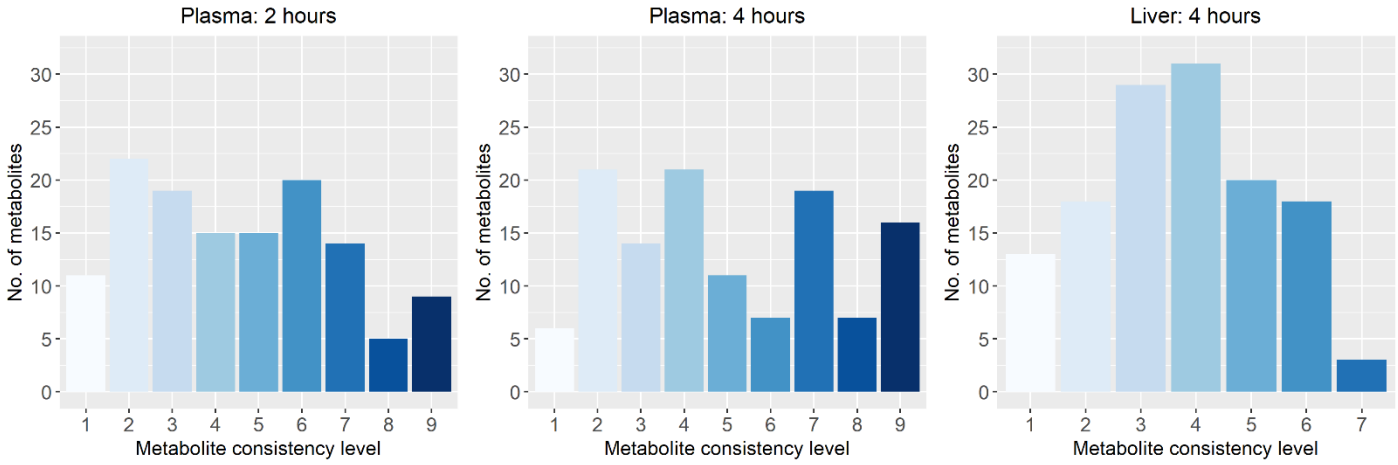


Figure 2.9. Bar charts to show the number of metabolites for each biomarker consistency level based on the point criteria described in *Table 2.1*. A higher consistency level indicates significance across multiple methods.

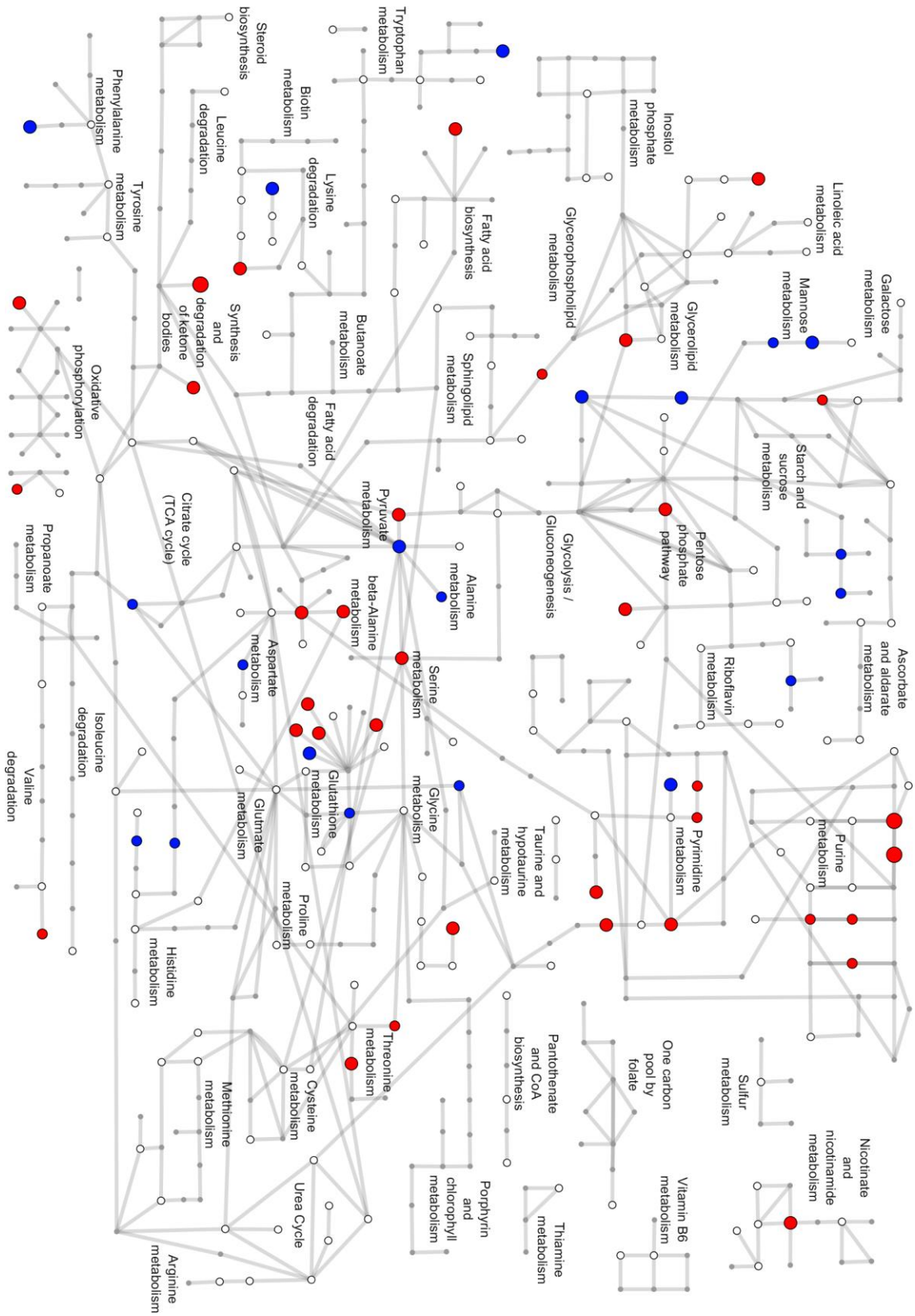


Figure 2.10. A network heatmap representation of the liver significant metabolite subset, nodes as metabolites and edges as reactions. Metabolites in red are increased in the low dose rats compared to control, blue are decreased whilst white are those found not significantly different. Metabolites not measured at all in this study were coloured in grey. Node size is proportional to their consistency level.

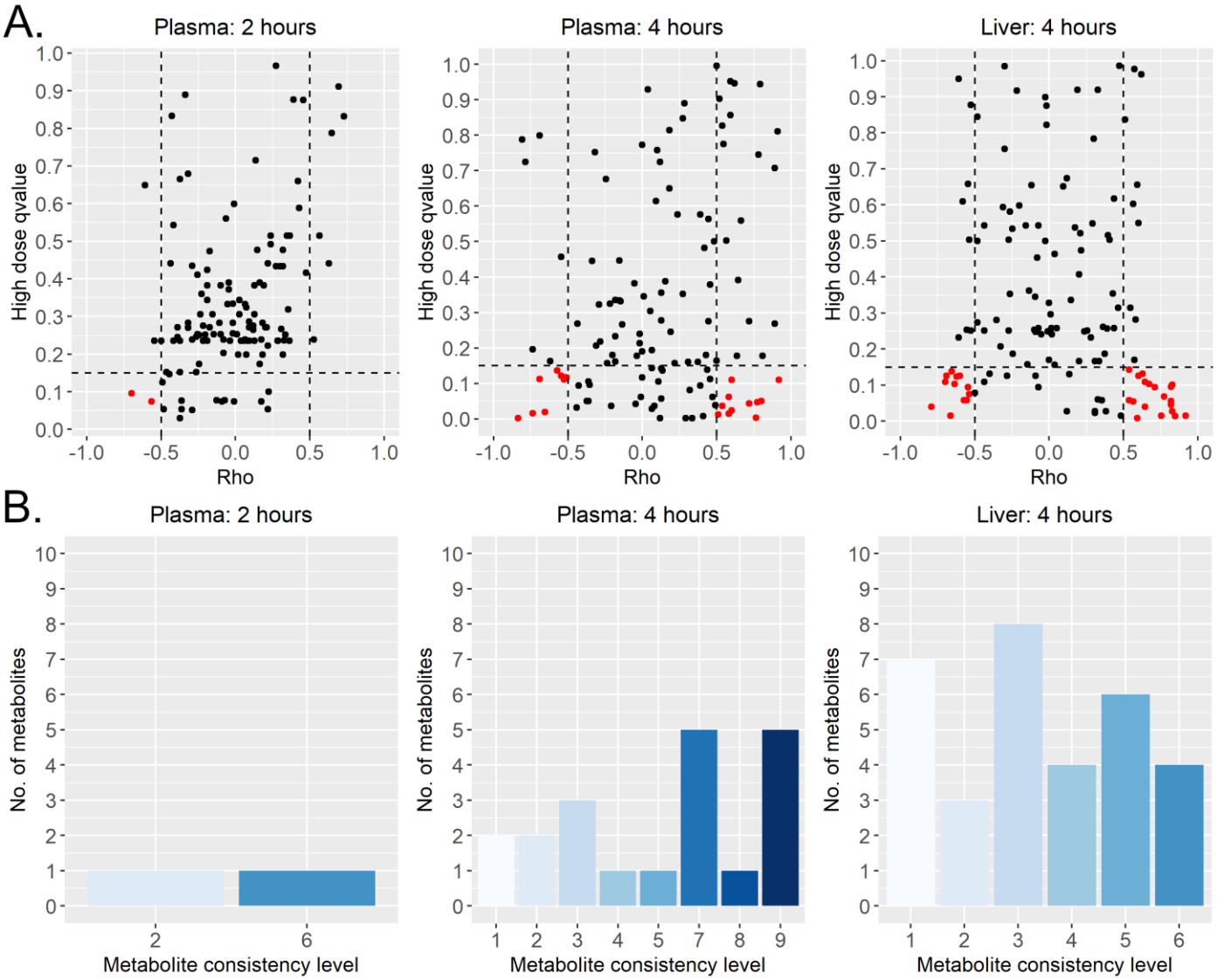


Figure 2.11. A) Scatter plots of a metabolite's control vs high dose q-value against its correlation coefficient Rho, which indicates the correlation between the level of the metabolite in a sample and the drug exposure level of the sample. Metabolites in red meet the criteria for potential biomarkers. **B)** Bar charts of the control vs. low dose consistency levels of the metabolites which meet the criteria for potential biomarkers**.**

2.4. Discussion

2.4.1. Tissue dose response

In multi-tissue organisms, each tissue has different metabolic activity based on its role, function and environment. In animals, the highly metabolically active tissues, like in the brain and liver, also have relatively high energy demands. This makes these tissues particularly susceptible to mitochondrial toxins, and the reason why symptoms related to these organs, such as hyperammonemia and neurodegeneration, are some of the most common clinical signs of mitochondrial dysfunction. This study aims to shed light on the metabolic adaptations which occur in a high energy organ, the liver, during mitochondrial dysfunction.

The low dose administered to the rats was known to be non-lethal. It was therefore not unexpected that all three of the analyses performed did not show a large significant difference between the low dosed and control rats (*Figure 2.3*). The minor differences were the metabolic responses due to the mild mitochondrial dysfunction which were able to cope with the added stress. These responses, despite being relatively minor, are of key interest as they could be used to identify mitochondrial dysfunction early and shed some insight into the order in which metabolic adaptions occur. These reasons ensured the investigation into the low dose liver sample despite showing only a minor significant difference between dosed and control. The higher dose, known to be lethal if the animals were left for greater than four hours, was identified as significantly different between dosed and control in both the PCA plot and hypothesis testing. The higher level of mitochondrial dysfunction caused huge amounts of metabolic adaptations which were unable to cope with the added stress, a likely situation in multiple organs and the major reason for the dose lethality.

The Z-score distribution for the high dose liver samples shows that the metabolite levels in the liver do not seem to be an accurate reflection of an organism’s drug exposure level and, by extension, mitochondrial dysfunction level (*Figure 2.1, 2.2*). Despite having a largely variable drug exposure level in each sample, particularly sample 17, the variance of the Z-scores for each sample does not differ greatly within the high dose group. Along with the low significance of the liver low dose

samples, this indicates that the adaptations which occur at all levels of mitochondrional dysfunction only caused subtle changes in metabolite concentrations in the liver, which is most likely caused by the essential functions of the tissue, such as converting ammonia into urea, restricting the adaptability of the tissue.

Blood plasma transports many compounds around an organism's body, including metabolites, hormones, waste products and proteins, and contains no mitochondria. Therefore, the concentration of metabolites in the plasma reflects the organism metabolic state, a cumulative representation of every organ's metabolic activity. Because of this, and the fact that it is easily accessible and in abundance, blood plasma is commonly used to monitor diseases such as diabetes. Based on the results of this study, blood plasma changes are seen during mitochondrial function which indicates that blood plasma can also be used to monitor both mild and fatal levels of mitochondrial dysfunction once a biomarker has been identified.

The distribution of Z-scores for all the dosed rats in the plasma showed a better reflection of each rat's drug exposure level (*Figure 2.2*). The high dose group had a generally larger variance than the low dose group. The high dose group also had a larger within group variance which reflected each samples exposure level. This is most likely because the plasma metabolite levels reflect the cumulative subtle adaptations occurring in each of the organism's organs during mitochondrial dysfunction which are easier to identify in the metabolite concentrations than the subtle changes occurring in the liver. These results show that plasma is a viable body fluid for detecting and monitoring mitochondrial dysfunction. In addition, based on the PCA plot of the plasma four hour samples, it should be possible to distinguish between mild and fatal levels of mitochondrial toxicity (*Figure 2.3*).

The low dose rats had significantly different plasma metabolite levels two hours after dosing. The delay of two hours can be attributed to many things such as the time needed for the transportation and accumulation of the drug in the high energy organs. Rats belonging to the high dose group showed significantly different plasma profiles potentially as early as 0.5 hours after dosing, based on the hypothesis testing (*Figure 2.4*). The difference in response time of the low dose compared to high dose suggests that the response time of an organism's metabolic activity to a

known mitotoxin has the potential to be used as an indicator for its toxicity level, and that early measuring of biomarkers may be necessary but further exploration of this theory is required.

2.4.2. Metabolic adaptations to mild mitochondrial dysfunction

The metabolomics data set generated for this study measured 844 metabolites in each plasma sample, over five time points, and 787 metabolites in the terminal liver sample which, in comparison to most metabolomics studies, would be considered a large and comprehensive data set. However, due to the huge size of the human metabolome and its dynamic nature, the interpretation of metabolomics data is difficult and can be ambiguous. The measuring of a metabolic profile at a single point in time acts as a snapshot of an organism's metabolic activity. Metabolic profiles are extremely dynamic so multiple snapshots would be required to accurately identify the precise metabolic changes occurring over time. In addition, changes in metabolite concentrations between case and control do not have a clearly defined translation into changes in pathway activity between case and control. For example, a metabolite which is increased in case compared to control could be due to either an increase in upstream pathway activity or a decrease in downstream pathway activity. Therefore, the results of the analysis performed on this data set should be considered only the first step in elucidating the metabolic adaptations occurring in mild mitochondrial dysfunction.

The significant metabolite subsets for each time point contained around 15% of the total measured metabolites. This subset could have been expanded, and a larger set would allow for better interpretation for metabolic activity, by selecting a more lenient VIP score threshold. However, this would be a trade off with potentially adding many false positives which would only add greater uncertainty to an already difficult situation to interpret (*Figure 2.8*). The subsets are large enough to enable the identification of pathways altered in the low dose rats compared to controls. The potential reasoning for the pathway changes, along with any problems associated with the altered activity, can be hypothesised and the problems associated with the perturbation discussed. The complete significant metabolite subsets for each

sample, classified into their primary pathways, can be found in *Supplementary File 2.3*. A small portion of each of the subsets contained metabolites which are either unknown (labelled as 'X') or metabolites which are externally related, such as those related to diet. These metabolites provided no insight into the metabolic adaptations occurring and were subsequently ignored.

Current understanding of mitochondrial function and the network of pathways involved in oxidative phosphorylation (OXPHOS), the main pathway for ATP generation in mitochondria, leads to expected perturbations in certain pathways during mitochondrial dysfunction. The inhibition of OXPHOS causes a lack of ATP production which must be compensated for by the other major ATP producing pathways, which includes glycolysis, fatty acid β-oxidation and amino acid catabolism (*Figure 2.12*). Each of these pathways generates ATP by feeding the tricarboxylic acid (TCA) cycle and are expected to have increased activity during mitochondrial dysfunction. A second major effect of OXPHOS dysfunction is on the NAD/NADH ratio as NADH builds ups. Changes in the NAD/NADH ratio can have widespread metabolic implications as it impacts many enzymes dependent upon NAD or NADH, e.g. dehydrogenases.

Glycolytic metabolism

Glycolysis is the conversion of sugars, such as glucose or fructose, into pyruvate which generates two ATP molecules per sugar and feeds the TCA cycle via acetyl-CoA. The pathway is considered a faster, but less efficient production of ATP than OXPHOS and is expected to have increased activity during mitochondrial dysfunction. The liver significant subset contained 12 different metabolites involved in glycolysis, including a decrease in multiple sugars such as glucose, fructose and mannose, and an increase in the downstream metabolite phosphoenolpyruvate (PEP). These changes are consistent with an increase in glycolysis in the liver during mitochondrial dysfunction. Both plasma subsets contained only two metabolites involved in glycolysis, with only mannose being part of both subsets. The lack of glycolytic metabolites in the plasma indicates that increased glycolysis may not be detectable in the plasma during mild mitochondrial dysfunction. Mannose, a sugar monomer which can be converted into glucose, was increased at both plasma time points, the opposite to its measurement in liver. Mannose metabolism is heavily

involved in glycosylation, a post-translation modification. A system under increased stress such as mitochondrial dysfunction will undergo many adaptations which could potentially require an increased amount of post-translational modifications, a potential reason for the increased levels of mannose in plasma during mitochondrial dysfunction.

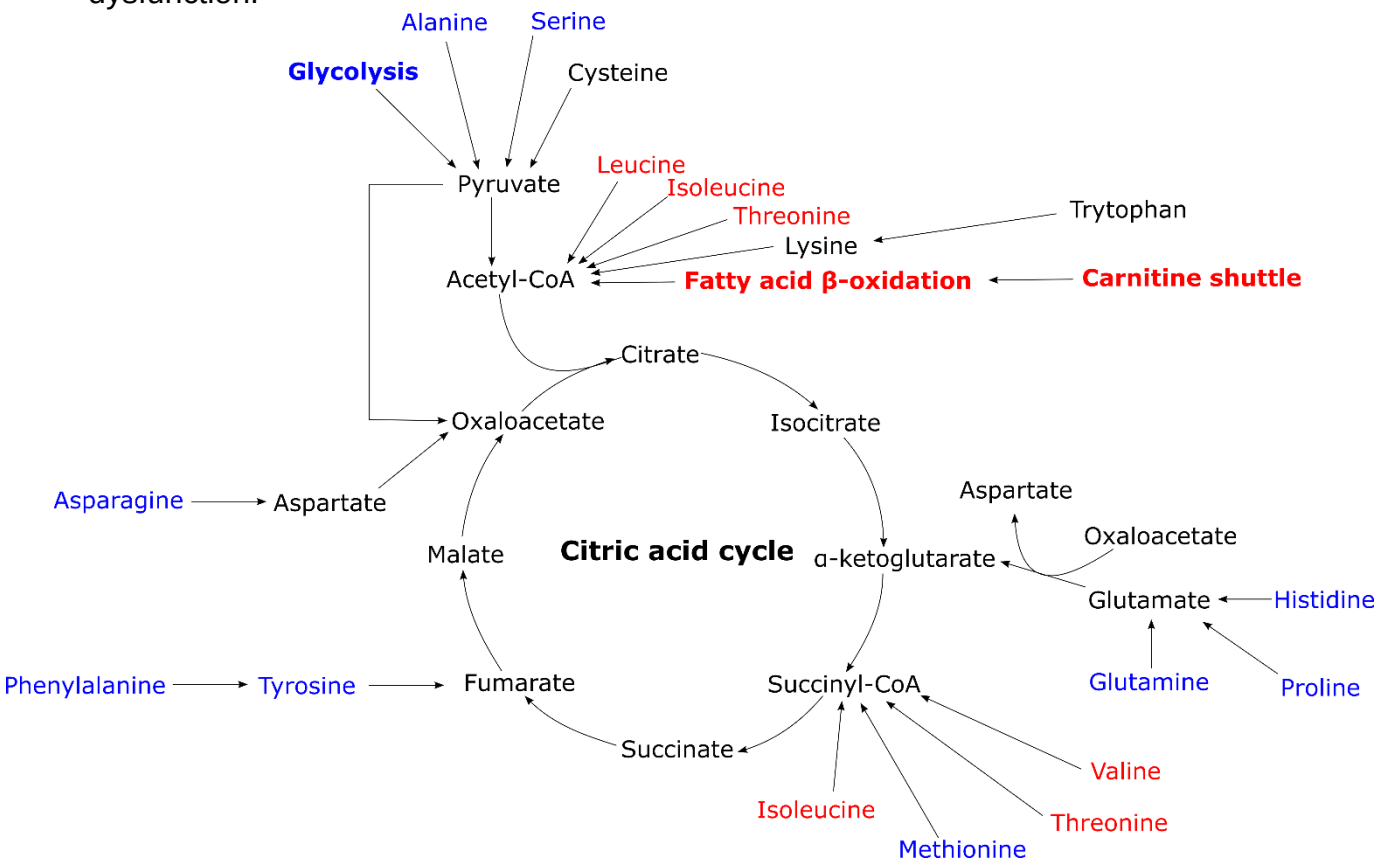


Figure 2.12. Overview of the citric acid cycle and the interactions with glycolysis, fatty acid β-oxidation and various amino acid metabolism pathways. Red pathways have increased levels of intermediates in the plasma and liver during mild mitochondrial dysfunction whilst blue pathways have decreased levels.

Fatty acid β-oxidation

Fatty acid β-oxidation is the breakdown of fatty acids into acetyl-CoA to feed the TCA cycle. Although no ATP is directly produced during this process, the pathway is expected to have increased activity during mitochondrial dysfunction to feed the TCA cycle and to regenerate OXPHOS electron donors, NADH and $FADH_2$, from their reduced forms. All three time point subsets had many metabolites which are involved in β-oxidation, the plasma 2 hour subset had 20 metabolites, the plasma 4 hour subset had 11 metabolites and the liver 4 hour subset had 7 metabolites

(*Supplementary File 2.3*). Most of these metabolites were fatty acid chains of varying length, all bar one of which were found to be increased at all time points. This is consistent with a large increase in the mobility of fatty acids and an increase in fatty acid β-oxidation in the liver, behaviour which is expected to occur in most tissues.

The large representation of fatty acids in the plasma 2 hour subset suggests that an increase in fatty acid β-oxidation is one of the immediate responses to mitochondrial dysfunction. The smaller set of fatty acids in the plasma 4 hour subset suggests that either fatty acid β-oxidation decreases in activity over time or that more extreme adaptions occur at later stages of the adaptation process causing other metabolites to become more significantly different. The plasma 4 hour subset contained the increase of three different 3-hydroxy fatty acids, one of which was also present and increased in the plasma 2 hour subset. The accumulation of 3-hydroxy fatty acids in plasma is indicative of defects in fatty acid β-oxidation and has been identified in patients suffering from glycogen storage disease [102] and shown to mimic long-chain 3-hydroxyacyl-CoA dehydrogenase deficiency [103], an error in the metabolism of long-chain fatty acids. This indicates that erroneous fatty acid β-oxidation was occurring somewhere within the organism which may be the reason for the lower representation of fatty acids in the plasma 4 hour subset compared to the plasma 2 hour subset.

The transportation of long-chain fatty acids across the inner mitochondrial membrane for fatty acid β-oxidation requires the carnitine shuttle (*Figure 2.13*). Fatty acids are converted into carnitine-fatty acid complexes in the cytosol, which freely move across the outer mitochondrial membrane and get transported across the inner mitochondrial membrane by carnitine-acylcarnitine translocase (CACT). The complex is then separated into a fatty acid and carnitine in the matrix, where the fatty acid enters β-oxidation. Both plasma subsets contained around ten different fatty acid carnitines, almost all of which were increased, along with a decrease in carnitine. The accumulation of fatty acid carnitines has been identified in the blood plasma of patients with fatty acid and branched-chain amino acid disorders [104]. Fatty acid carnitines that accumulate in the plasma are a spill over from fatty acid transport to tissues, where different tissues have been shown to produce a different fatty acid carnitine plasma profile [105]. As the inner mitochondrial membrane is

impermeable, an accumulation of fatty acid carnitines in the plasma is most likely the result of a defect in CACT, the transport across the inner membrane. In patients with CACT deficiency, the accumulation of fatty acid carnitine does occur in the plasma and has a high mortality rate [106]. The decrease in carnitine supports the argument of a transport defect as carnitine is recycled through the carnitine shuttle process, and a blockage in this process would result in carnitine being trapped in the fatty acid carnitine complex. The liver subset contained four different carnitine-fatty acids, three of which were decreased, which does not clearly indicate that the liver is the tissue encountering the carnitine shuttle problem.

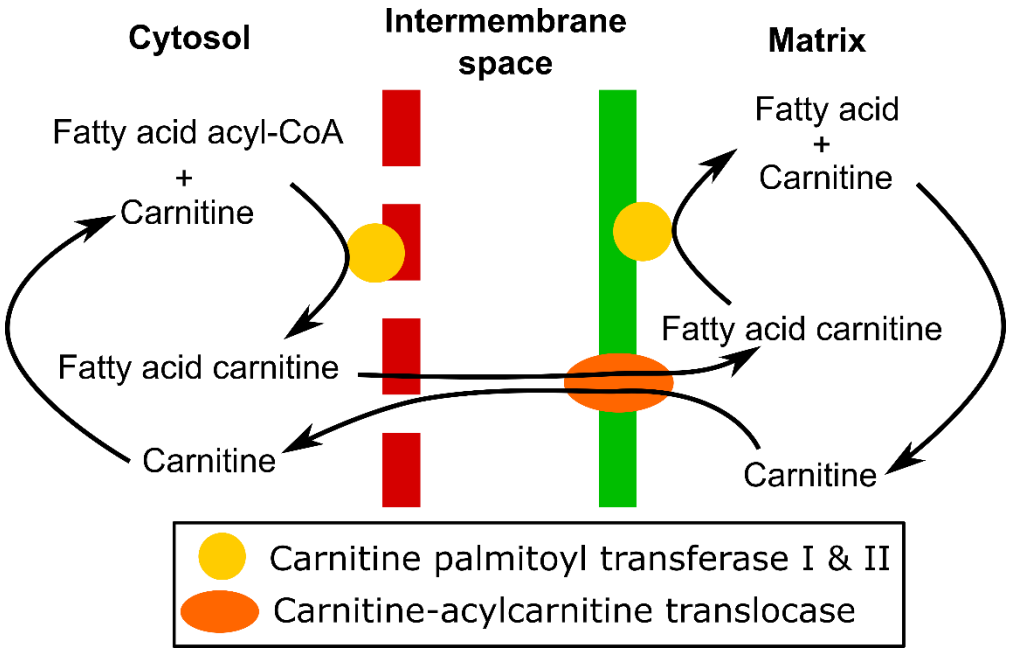


Figure 2.13. Overview of the carnitine shuttle, the method of transporting fatty acids from the cytosol into the mitochondrial matrix for fatty acid β-oxidation.

Amino acid metabolism

Various amino acids can be metabolised and fed into the TCA cycle (*Figure 2.12*). The liver subset contained three amino acids which were increased: serine, threonine and valine, while alanine and glutamine were decreased. The decrease in alanine and glutamine is most likely related to their intricate involvement in the liver's essential functions. Alanine can be directly converted into pyruvate using alanine transaminase (ALT), an enzyme which is regularly monitored in patients to test for liver function. During mitochondrial dysfunction, an increase in alanine metabolism

will help feed the TCA cycle for ATP production. However, a common hallmark in patients with mitochondrial disease is an increase in alanine concentration in the plasma, where alanine is produced from pyruvate in the liver and exported to the plasma [107]. The contradicting result identified in this study suggests that the excretion of alanine, which also comes with lactate and not seen in this study, is an adaptation which may only occur during high levels of mitochondrial dysfunction.

A potential reason for the glutamine deficiency is that glutamine has been implicated in maintaining an organism's acid-base balance [108], where it is released from the liver and collected in the kidney during ketosis, the accumulation of ketone bodies in the blood. Ketone bodies are molecules produced in the liver from fatty acids which are then transported to other organs, particularly the heart and brain, and metabolised as an energy source. All three of the subsets had increased level of 3-hydroxybutyrate, a ketone body, and acetoacetate was increased in the plasma 2 hour subset, suggesting that ketosis was present during mild mitochondrial dysfunction. The decrease in glutamine could also be due to increased glutaminolysis metabolism, which is a critical pathway for cells undergoing high levels of glycolysis, which is expected during mitochondrial dysfunction [109]. Glutamine metabolism was found to be sufficient in maintaining the levels of acetyl-CoA and oxaloacetate when the mitochondrial pyruvate transporter was inhibited [110], making it a relatively efficient pathway for feeding the TCA cycle.

Serine and threonine are two structurally and functionally similar amino acids, and their accumulation in liver could be due to many different reasons (*Figure 2.14*). Serine activity has recently gained much attention in cancer research, where it has been established as an important amino acid for cell growth and proliferation [111]. It is also increased as part of the remodelling of one-carbon metabolism that occurs during mitochondrial dysfunction [112]. The metabolism of serine and threonine in the liver is predominantly via their conversion into glycine [113,114]. The conversion of serine into glycine produces pyruvate, which can feed the TCA cycle, and the produced glycine is used to regenerate NADH from NAD+ in the mitochondria via the glycine cleavage system [115]. The NAD/NADH ratio in mitochondria is critical for mitochondrial function [116]. The liver subset identified an increase in NAD+, indicating a shift in the NAD/NADH ratio and supports the argument that both serine

and threonine metabolism, in addition to the glycine cleavage system, are important during mitochondrial dysfunction.

Serine also plays a key role in one-carbon metabolism by acting as a one-carbon donor for the tetrahydrofolate (THF) cycle [111], which supports purine and pyrimidine nucleoside synthesis. The production of purine and pyrimidine nucleosides is necessary for protein synthesis [117]. During times of cellular stress, such as mitochondrial dysfunction, many changes in pathway activity are expected and requires a large amount of protein synthesis, e.g. increased production of glycolytic enzymes and proteins to counter increased oxidative stress. The liver significant metabolite subset contained many nucleosides including guanosine and cytidine, which were increased, along with an increase in multiple other forms of nucleoside such as 7-methylguanine and pseudouridine. This accumulation of nucleosides supports the hypothesis of increased protein synthesis during mitochondrial dysfunction.

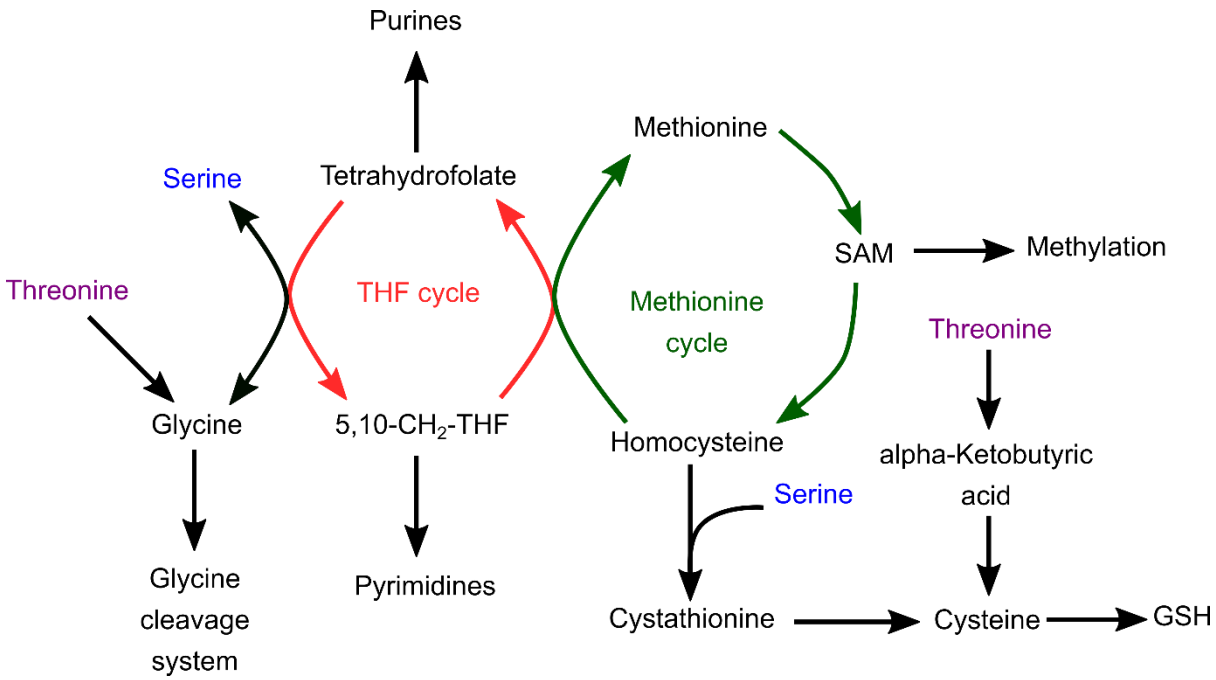


Figure 2.14. Overview of serine and threonine metabolism and their involvement in the tetrahydrofolate (THF) cycle, the methionine cycle and the production of glutathione (GSH).

The THF cycle also supports methionine metabolism and the production of *S*-adenosyl methionine (SAM), which is critical for protein methylation, a post-translational modification. If mitochondrial dysfunction causes an increase in protein synthesis, this behaviour will most likely accompany an increase in post-translation modifications. The same discussion can be made to rationalise the increase in threonine as it has been implicated in the methylation of histones which play an important role in gene regulation [118], closely related to protein synthesis.

Nicotinamide adenine dinucleotide phosphate (NADP+) in its reduced form NADPH, can be used in replacement of NADH for certain critical mitochondrial reactions and can be converted directly into NADH [119]. The THF cycle has been identified as a major regenerator of NADPH from NADP+ [116]. NADPH also has its own critical function in maintaining glutathione metabolism [120], important for reducing reactive oxygen species (ROS) levels. Minor levels of ROS are produced during normal OXPHOS [121], but during mitochondrial dysfunction the levels of ROS are known to increase massively due to electron leak [122], which has fatal effects on mitochondria, e.g. mtDNA damage. Glutathione (GSH) is one of the most important scavengers of ROS, by reacting with free ROS it is converted into glutathione disulphide which can be regenerated back into GSH using NADPH. The ratio of GSH/GSSG is used as a marker for cellular oxidative stress levels [123].

In the liver significant metabolite subset, GSH is decreased along with an increase in multiple γ-glutamyl amino acids. An important metabolite in the cyclical synthesis of GSH is 5-oxoproline, which is primarily generated by the degradation of GSH (*Figure 2.15*). However, 5-oxoproline can also be generated from the degradation of γ-glutamyl amino acids, generating 5-oxoproline and releasing the amino acid. This process has been suggested to occur when cysteine levels are limited [124], as cysteine is also an important metabolite in the cyclical synthesis of GSH. Multiple γ-glutamyl amino acids were decreased in both plasma subsets, with 5-oxoproline found decreased at plasma 4 hours which suggests that this activity was occurring organism wide resulting in the mass transportation of γ-glutamyl amino acids around the organism. Four γ-glutamyl amino acids were found increased in the liver whilst γ-glutamyl valine was found to be decreased along with an increase in valine. This supports the theorem of increased γ-glutamyl amino acid degradation and identifies a

potential γ-glutamyl amino acid degradation affinity. These facts confirm that during mild mitochondrial dysfunction there is a high level of oxidative stress, and that the regeneration of NADPH is a critical role of the THF cycle. Furthermore, both serine and threonine can be metabolised into cysteine, required for GSH synthesis, via cystathionine and α-ketobutyric acid respectively, the latter of which was found increased in plasma 4 hours. Therefore, serine and threonine accumulation in liver may be related to their involvement in maintaining redox homeostasis during mild mitochondrial dysfunction.

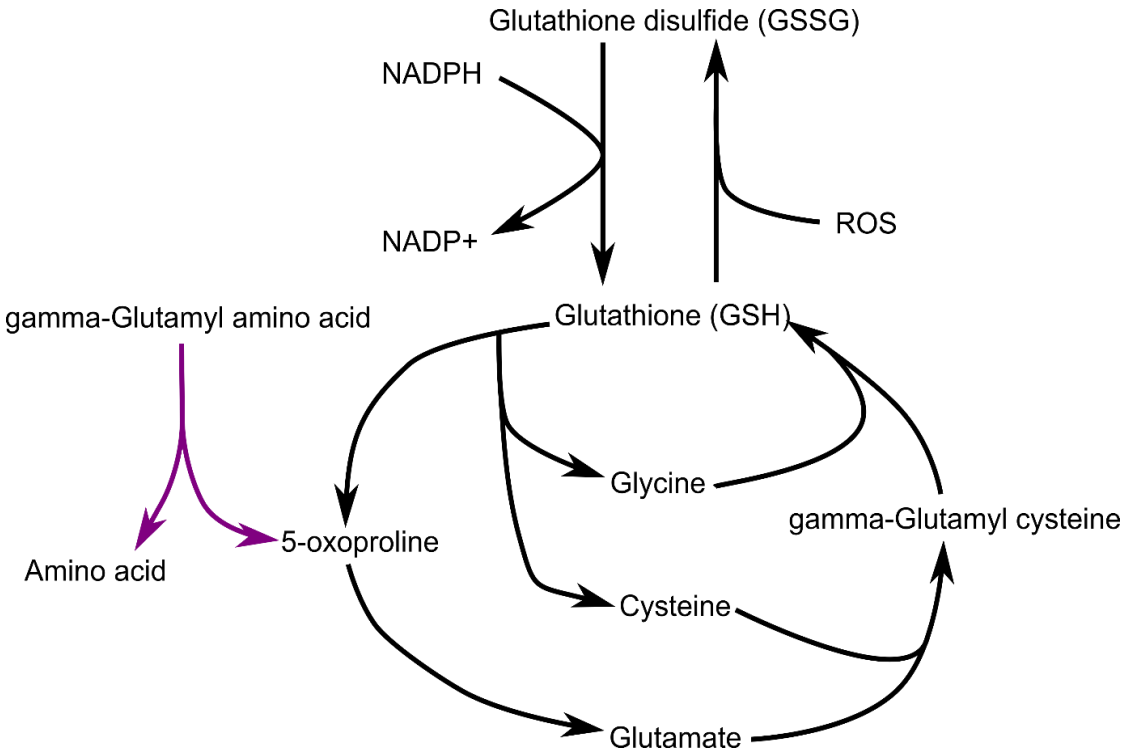


Figure 2.15. Overview of glutathione metabolism, biosynthesis and regeneration from glutathione disulfide. Purple shows the compensatory mechanism of γ-glutamyl amino acid degradation.

The availability of serine in mitochondria has been implicated in mitochondrial dynamics [125]. Serine is a critical metabolite in the production of phosphatidylserine (PS). PS, along with phosphatidylcholine (PC) and phosphatidylethanolamine (PE), are major components of both mitochondrial membranes [126]. The accumulation of fatty acids, which this study has confirmed to be occurring during mild mitochondrial dysfunction, has been shown to promote the synthesis of phospholipids [127,128]. Phosphatidylserine is produced by the reaction of PC or PE with serine, which are

themselves produced by choline and ethanolamine. In the liver metabolite subset, both choline and CDP-ethanolamine, an intermediate metabolite in PE synthesis, are both increased. This evidence suggests that the production of phospholipids occurs during mitochondrial dysfunction, most likely as part of the fission/fusion repair mechanism in which serine and choline both play a key part which adds to the list of potential reasons for the identified serine accumulation. The important functions of serine and threonine and their increased levels highlights them both as important metabolites to be studied further in the context of mitochondrial dysfunction.

Phospholipids

Glycerophospholipids are glycerol-based phospholipids, the major component of cell membranes [129]. The breakdown of cell membranes occurs during neurodegeneration, a common symptom of mitochondrial dysfunction, resulting in the accumulation of choline containing metabolites in the brain [130]. All three subsets contained a large amount of different glycerophospholipids, with the liver having a decreased level of glycerophosphorylcholine and glycerophosphorylethanolamine, along with an increase in glycerol-3-phosphate, a major component of glycerophospholipids. These indicate the breakdown of cell membrane proteins in the liver, with the plasma shuttling glycerophospholipids around the body to be potentially broken down in other tissues. The breakdown of glycerophospholipids in the liver can be used as a source of choline via the Kennedy pathway [131] and ultimately leads to the production of PC [132]. An alteration in choline metabolism has been identified as a hallmark of cancer [133], with choline metabolites and PC being integral for cell growth and death [134]. This highlights the importance of choline metabolism during mitochondrial dysfunction. Moreover, the breakdown of glycerophospholipids suggests that serious structural changes to the cellular membrane occurs during mild mitochondrial dysfunction.

Branched chain amino acids

Valine, isoleucine and leucine are the branched-chain amino acids (BCAA's), whose metabolism feeds directly into the TCA cycle. BCAA metabolism is one of the amino acid metabolism pathways which is activated during exercise [135] and thought to be promoted by fatty acid β-oxidation [136], which suggests that BCAA metabolism activity is expected to increase during mitochondrial dysfunction. The accumulation of BCAA's has been linked to the inhibition of glycolysis in the heart [137]. Valine was identified as increased in the liver, but there was also many other intermediate BCAA metabolites which were found increased in both the liver and plasma. In this study, the intermediates of BCAA metabolism did not seem to cause inhibition of glycolysis. However, almost all the BCAA metabolism intermediates identified in all three subsets are biomarkers for a range of different metabolic defects.

Both 4 hour subsets had 3-hydroxy-2-ethylpropionate increased, the accumulation of this metabolite is the hallmark of short/branched-chain acyl-CoA dehydrogenase deficiency [138], a defect in isoleucine metabolism. All the subsets also had an increase in 3-hydroxy-3-methylglutarate which is only produced by defective 3-hydroxy-3-methylglutaryl-CoA lyase [139], the final step in leucine degradation which catalyses the conversion of 3-hydroxy-3-methylglutaryl-CoA to acetyl-CoA and acetoacetate. Finally, both plasma subsets had an increase in 3-methyl-2-oxobutyrate, which is only produced by incomplete BCAA metabolism and a hallmark of maple syrup urine disease [140], and 3-hydroxyisobutyrate, which is an intermediate of valine metabolism. The accumulation of 3-hydroxyisobutyrate is caused by 3-hydroxyisobutyryl-CoA dehydrogenase deficiency and has been identified as a complex I-III inhibitor [141]. These indicate a defect in all three BCAA metabolism pathways which is related to multiple enzymes. The cause of each of the defects is not obvious, but each of the defects mentioned causes serious clinical symptoms. Therefore, the identification of the cause of the suspected defects should be the focus of future study.

Adenine nucleotides

The decrease in ATP production caused by mitochondrial dysfunction is expected to cause a decrease in the ATP/ADP ratio. In the mitochondria, two ADP molecules can be recycled into a single ATP molecule, and one adenosine monophosphate (AMP). Evidence of this recycling was present in the liver subset, with AMP increased along with its degradation products: adenylosuccinate, xanthine and xanthosine [142]. However, in the plasma 4 hour subset there was an increase in succinyladenosine (SAdo) which is only present in the plasma of patients with adenylosuccinate lyase (ADSL) deficiency [143]. Therefore, this recycling process is prevalent during mild mitochondrial dysfunction, and could also have erroneous activity, suggesting that SAdo could be an important biomarker and should be studied further.

2.4.3. Mitochondrial dysfunction biomarkers

The analysis of the metabolomics data set performed in this study identified a large set of metabolites which were highly significantly different between the low dose and control rats, all of which could be potential biomarkers for mild mitochondrial dysfunction. Using a consistency level threshold of six for plasma and five for liver, each of the three sets identified around 40 highly significant metabolites. Each subset consists of metabolites which are common among all three time points along with metabolites which are unique to a single time point. However, the biological relevance of the metabolite must be taken into consideration if they are to be labelled as biomarkers for mild mitochondrial dysfunction. *Table 2.3* shows the biologically relevant biomarkers for mild mitochondrial dysfunction for each time point using the previously discussed adaptations as the basis for selection, along with some previously unmentioned metabolites which were included based on their extremely high significance. The plasma biomarkers identified are the most important due to the availability of blood plasma during experimentation. However, the liver biomarkers have been included in the table as they may be of relevance for future exploratory studies and many of them provide supporting arguments for the inclusion of many of the plasma biomarkers.

The unique biomarkers for the plasma 2 hour time point are a host of different fatty acids and two different ketone bodies, all of which were increased in the low dose rats compared to controls. The accumulation of fatty acids and ketone bodies in the plasma could be used as an identifier of early stage mild mitochondrial dysfunction but this result needs to be verified by future studies. The animals dosed in this study were not fasted before dosing, but the effects of fasting can cause an increase in fatty acid β-oxidation and ketosis. Future studies should ensure the same conditions are used to ensure that this behaviour is a response to the mitochondrial dysfunction and not animal fasting.

The accumulation of 3-hydroxy fatty acids was also a unique biomarker for the plasma 4 hour subset. This is a biomarker of an error in fatty acid metabolism which seems to be present during late stage mild mitochondrial dysfunction. The accumulation of this metabolite could be used as a biomarker for late stage mitochondrial dysfunction. The amino acids glutamine and citrulline were uniquely decreased in the plasma 4 hour subset and can be used as late stage mild mitochondrial dysfunction biomarkers. Glutamine has many important functions which were discussed in earlier sections, whilst citrulline participates in the urea cycle. Late stage mild mitochondrial dysfunction also has biomarkers related to the TCA cycle, citrate and nicotinamide, an important structural component of NADH. The accumulation of these metabolites is indicative of the TCA cycle blockages and general struggles of the mitochondria to keep the TCA cycle running efficiently whilst OXPHOS is inhibited. Finally, the accumulation of succinyladenosine, a metabolite which is only present in human plasma when there is an issue in the purine nucleotide cycle, is present during late stage mild mitochondrial dysfunction, making it a potentially important biomarker.

There were many metabolites which were highly significant in both plasma subsets, most of these have already been discussed in previous sections but the following are of interest as mild mitochondrial dysfunction biomarkers. Mannose is a sugar which would be expected to be decreased due to its consumption for glycolysis in organs. Therefore, its accumulation makes it an interesting biomarker for mild mitochondrial dysfunction, potentially related to its involvement in glycosylation. There are three amino acids which were decreased in both plasma subsets and could be used as

biomarkers for mild mitochondrial dysfunction: alanine, proline and methionine. Alanine is important for pyruvate production to feed glycolysis, methionine for its methionine cycle relating to methylation and GSH production, and proline for its metabolism which has recently emerged as an important pathway in cancer [144].

Based on this study, mild mitochondrial dysfunction seems to cause problems with BCAA metabolism. 3-Hydroxyisobutyrate, a biomarker of defective isoleucine metabolism and a potential inhibitor of complex I-III, seems to accumulate in both plasma subsets making it a potentially critical biomarker for mild mitochondrial dysfunction. Mild mitochondrial dysfunction also causes a large amount of oxidative stress. The decrease in multiple γ-glutamyl amino acids was seen in both plasma subsets, and their accumulation seen in liver. These metabolites are therefore particularly important during mild mitochondrial dysfunction and could be used as biomarkers. In this study, γ-glutamyl alanine and γ-glutamyl methione were identified as high significant biomarkers. Finally, a decrease in mevalonate was also identified in both plasma subsets. Mevalonate is part of the mevalonate pathway in which isoprenoids, such as heme and steroid hormones, are produced from 3-hydroxy-3-methylglutaryl-CoA, an intermediate in ketogenesis. The reason for the decrease in mevalonate is not obvious but may be related to the increased production of ketone bodies. Nonetheless, mevalonate was included in the table of biomarkers because of its extremely high significance and should be investigated in further studies.

The metabolites which were identified as potential biomarkers for all levels of mitochondrial dysfunction are shown in *Table 2.4*. These are the metabolites which are significantly different between case and control for both the low dose group and the high dose group, along with their measured metabolite level in each sample being correlated to the drug exposure level of the sample it was collected from, across all dosed samples. It should be noted that due to the low number of samples belonging to the high dose group, these results will need to be further experimentally tested to ensure greater confidence in the results.

Pathway	Plasma 2 hours	Plasma 4 hours	Liver 4 hours
Glycolysis	Mannose	Mannose	Maltopentaose, Maltotriose, Pyruvate
Fatty acid β-oxidation	Docosapentaenoate, Linoleate, Mead acid, Palmitate, Palmitoleate, Pentadecanoate	3-Hydroxylaurate, 3-Hydroxyoctanoate, 3-Hydroxysebacate	
Carnitine shuttle	Oleoylcarnitine	Carnitine, Oleoylcarnitine	Deoxycarnitine
Amino acid	Alanine, Proline. Methionine	Alanine, Asparagine, Glutamine, Proline, Citrulline, Methionine	Serine
BCAA metabolism	3-Hydroxyisobutyrate	3-Hydroxy-2-ethylpropionate, 3-Hydroxyisobutyrate	3-Hydroxy-2-ethylpropionate, 3-Hydroxy-3-methylglutarate
Phenylalanine metabolism	Phenyllactate	Phenyllactate	Phenyllactate

Dipeptide	Cyclo(leu-pro), Cyclo(L-phe-L-pro)	Cyclo(gly-pro), Cyclo(leu-pro), Cyclo(L-phe-L-pro)	
Ketone body	3-Hydroxybutyrate, Acetoacetate		
TCA Cycle		Citrate	NAD+, Pantethine
Purine metabolism		Succinyladenosine	Adenylosuccinate, AMP
Glycerophospholipid	1-Arachidonoyl-GPC, 2-Arachidonoyl-GPI	1-Linoleoyl-GPE, 1-Oleoyl-GPE	Glycerol 3-phosphate, Glycerophosphorylethanolamine
Choline metabolism			Choline
Oxidative stress	γ-Glutamyl alanine, γ-Glutamyl methionine	5-Oxoproline, γ-Glutamyl alanine, γ-Glutamyl methionine	Anserine, γ-Glutamyl glutamate, γ-Glutamyl tryptophan, γ-Glutamyl tyrosine
Other	Mevalonate	Mevalonate, Nicotinamide	

Table 2.3. Proposed biomarkers of mild mitochondrial dysfunction. Each metabolite has a high consistency level (≥ 6 for plasma, ≥ 5 for liver). Metabolites in red were found increased in low dose compared to controls, whilst blue were found decreased.

For the plasma 2 hour time point, only 12% of the metabolites which belonged to the low dose significant metabolites subset were also found to be significantly different between high dose and control (q-value < 0.15), whilst the plasma and liver 4 hour subsets had around 35%. This shows that at the earlier stages of mitochondrial dysfunction, the adaptations caused by mild mitochondrial dysfunction cause markedly different changes in metabolite concentrations than the adaptations which occur during high, fatal levels of mitochondrial dysfunction. Therefore, further study of high mitochondrial dysfunction may enable the identification of an early metabolic biomarker unique to high levels of mitochondrial dysfunction. Of the metabolites which were found to be significantly different in both the low and high dose samples in the liver, 70% of these had significant correlation with drug exposure level, whilst the plasma 4 hour had 43%. This difference was most likely caused by the fact that the liver adaptations are restricted by its essential functions, e.g. the urea cycle. The restricted adaptability of the tissue means that the pathway adaptations which occur during fatal levels of mitochondrial dysfunction are similar to those which occur during mild mitochondrial dysfunction, just at varied levels of pathway activity.

The metabolites which fit the criteria of potential biomarkers for all levels of mitochondrial dysfunction were filtered based on their biological relevance (*Table 2.4)*. The accumulation of one of the 3-hydroxy fatty acids, 3-hydroxyoctanoate, along with the decrease in carnitine are both potential biomarkers. These suggest that the problem in fatty acid β-oxidization also occurs in high levels of mitochondrial dysfunction. Another potential biomarker is the accumulation of 3-hydroxyisobutyrate in the plasma, indicating a persistent problem with BCAA metabolism during all levels of mitochondrial dysfunction. This further highlights the importance for the future study of the cause of the erroneous BCAA activity in the context of mitochondrial dysfunction. The breakdown of the cell membrane also seems to occur in high levels of mitochondrial dysfunction shown by decreased levels of GPE and GPC being a biomarker in liver, with the decreased level of 1-linoleoyl-GPE identified as a potential plasma biomarker. Other potential biomarkers include the decrease in mevalonte, the accumulation of niotinamide, and the accumulation of 2-hydroxybutyrate.

2.5. Conclusion

The analysis of the metabolomics data set in this study identified key metabolic adaptions which occur during mild mitochondrial dysfunction, some of which are seemingly erroneous. The further study of these pathways and the identification of the causes of the defects could potentially identify therapeutic targets for patients with mitochondrial diseases. Moreover, the analyses allowed the identification of potential biomarkers for both mild and high levels of mitochondrial dysfunction which could be used to improve the drug development process and help monitor patients undergoing any form of medication with suspected mitotoxicity.

Pathway	Plasma 4 hours	Liver 4 hours
Glycolysis		Maltotriose, Sorbitol
Fatty acid β-oxidation	3-Hydroxyoctanoate	
Carnitine shuttle	Carnitine	
Amino acid		Glutamine, Serine, Threonine
BCAA metabolism	3-Hydroxyisobutyrate	3-Hydroxy-2-ethylpropionate
Dipeptide	Cyclo(leu-pro), Cyclo(L-phe-L-pro)	Cyclo(leu-pro)
Ketone body		3-Hydroxybutyrate
Glycerophospholipid	1-Linoleoyl-GPE	Glycerophosphorylethanolamine, Glycerophosphorylcholine
Choline metabolism		Choline
Oxidative stress	2-Hydroxybutyrate	Glutathione
Other	Mevalonate, Nicotinamide	

Table 2.4. Proposed biomarkers of complete mitochondrial dysfunction. Metabolites in red were found increased in low dose compared to controls, whilst blue were found decreased.

Chapter 3

Modelling mitochondrial dysfunction

3.1. Introduction

3.1.1. Improving mitochondrial dysfunction biomarker identification

The analysis of the metabolomics data set in the previous chapter highlighted critical pathways and potential biomarkers for mild mitochondrial dysfunction. The data set contained many significantly different metabolites which were whittled down to a small set of potential biomarkers based on their biological relevance. However, this biological relevance was based on mostly theoretical reasoning as our overall understanding of mitochondrial dysfunction is limited. Despite being considered a comprehensive study, the data set was not large enough to enable absolute confidence in the identified biomarkers and highlighted critical problems which need to be addressed in future studies.

Firstly, mitochondrial metabolism is organ specific. The essential functions of each organ mean that the network of pathways involved in mitochondrial metabolism is slightly different for each organ. These essential functions, such as the generation of urea in the liver, cause a restriction on the adaptability of the organ to mitochondrial dysfunction suggesting that each organ would have a potentially unique metabolic profile during mitochondrial dysfunction. The plasma metabolite levels reflect the collective adaptations occurring in the whole organism. Therefore, to accurately identify a plasma biomarker for any level of mitochondrial dysfunction, we will need to have a better understanding of the adaptions occurring in many of the highly metabolically active organs such as the brain, liver and kidney. This issue could be addressed in future metabolomics studies by collecting terminal metabolite levels of

multiple organs, enabling the identification of both organ specific and organism wide adaptations to mitochondrial dysfunction.

Secondly, adaptations to mitochondrial dysfunction are not necessarily linear. With only a handful of plasma metabolite time points, and one liver time point, all behaviour in between must be extrapolated, usually with the assumption that the behaviour seen in one time point leads directly to the next, in a linear fashion. In reality, between two time points there could be a huge inflection point giving wildly different behaviour between the two measured time points which would invalidate any identified biomarkers in the early measured time point. Such an inflection point would be the perfect behaviour for biomarker identification but is currently impossible to identify due to the intricacy of the metabolic network and our lack of understanding of how adaptations change over time. This issue could be negated by including many time points which are measured in small time intervals, allowing for the monitoring of metabolite changes over time. However, this is realistically only possible for plasma samples *in vivo*.

Thirdly, pathway activity inference from metabolite levels is difficult, but critical for biomarker identification. Ideally, the metabolites identified as biomarkers would be those which are involved in critical pathways specifically related to mitochondrial dysfunction, and not pathways which are just general organism stress responses. Most pathways involve hundreds of different metabolites, and each metabolite is generally involved in multiple different pathways. Therefore, the measured increase or decrease of a metabolite in a case sample compared to a control does not translate perfectly to an increase or decrease in the activity of a specific pathway, hindering the identification of mitochondrial dysfunction biomarkers. By measuring a much larger subset of metabolites, it would provide a much larger coverage of each of the pathways involved in mitochondrial metabolism and help identify the correct reason for the significant difference of a metabolite between case and control, allowing for better biomarker identification.

Fourthly, different types of mitochondrial dysfunction will most likely cause different metabolic adaptations. The mitochondrial dysfunction analysed in the previous chapter was caused by mitochondrial complex III inhibition, but mitochondrial dysfunction can be caused by defects in multiple different types of protein including

transporters, enzymes and the other mitochondrial complexes. Each of these would disrupt a different part of mitochondrial metabolism, resulting in different adaptations. The biomarkers identified in the previous chapter were labelled as mitochondrial dysfunction biomarkers, although it's more than likely that some of these may be mitochondrial complex III inhibition specific. For both patient monitoring and drug development it would be beneficial to have both general mitochondrial dysfunction biomarkers and biomarkers for specific types of mitochondrial dysfunction. Knowing the type of dysfunction occurring would enable measures to be taken to counteract the specific inhibition. Current biomarkers for mitochondrial diseases are extremely similar which makes precise patient diagnosis a difficult and time-consuming process. For example, the primary biomarkers for mitochondrial disease caused by all four complex deficiencies are lactic acidosis [145–148], hypoglycaemia [149] and hyperammonaemia [150], unique biomarkers for each type of complex deficiency are currently unknown. To address this issue, future metabolomics studies should be carried out on various types of mitochondrial dysfunction. The results can then be compared to identify the general mitochondrial dysfunction biomarkers, along with biomarkers for specific types of inhibition.

Finally, different levels of mitochondrial dysfunction will also cause different metabolic adaptations. In the previous chapter, the small sample size of rats exposed to high levels of the drug prevented the identification of specific biomarkers for fatal levels of complex III inhibition, but the analyses highlighted the differences in early stage mitochondrial dysfunction at varying levels of inhibition. Early detection of the level of mitochondrial dysfunction is critical for improving both the drug development process, and patient monitoring. A perfect biomarker for mitochondrial dysfunction would not only detect mitochondrial dysfunction, but also be able to identify the specific level of mitochondrial dysfunction occurring. Future metabolomics studies should include a much wider range of drug exposure levels, with a suitable number of samples, so that the different metabolic adaptations occurring at each level of inhibition can be explored, and a consensus biomarker identified.

3.1.2. Flux balance analysis: the *in silico* alternative to *in vivo* studies

To improve the identification of mitochondrial dysfunction biomarkers using metabolomics will require a large investment of both time and money. An alternative to performing a large amount of investigative *in vivo* studies is to use *in silico* predictions to direct *in vivo* studies. Flux balance analysis (FBA) is a constraint-based modelling technique which has become extremely popular for studying biochemical networks and has been successfully used in the modelling of whole genome unicellular metabolism [151]. Having a model of whole organism mitochondrial metabolism and being able to use the model to simulate multiple types of mitochondrial dysfunction would drastically improve our understanding of the metabolic adaptions during mitochondrial dysfunction and instil higher confidence in biomarker selection.

FBA models are mathematical representations of a system of chemical reactions, based on their stoichiometry [152]. The stoichiometric coefficients of every reaction in the system are combined to form a single stoichiometric matrix, which itself acts as a constraint on the system. During the simulation process, the system is assumed to be at a steady state. This constrains the flow of metabolites through the system by ensuring that for every metabolite in the system, the total amount being produced in the system must be equal to the total amount being consumed. The flow of metabolites through a single reaction in the system is known as a reaction flux, with a flux representing the forward reaction being positive and the reverse reaction being negative. Additional constraints can be placed on each reaction in the form of an upper and lower bound, defining the allowed minimum and maximum fluxes through a single reaction. For example, a reaction which is irreversible will have its lower bound set to zero, preventing the flow of metabolites in the reverse direction. The other essential component of an FBA model is its defining phenotype, known as the objective function. The objective function is the overall biological objective of the system which, in most biological networks, is the production of biomass for growth. In previous FBA models of mitochondrial metabolism, the objective function has been set to the production of ATP [153]. The combination of a steady state stochiometric matrix and the objective function creates a system of linear equations which can be solved using linear programming [154]. The result produces a list of

reaction fluxes which satisfies the optimization of the objective function which, in the case of mitochondrial metabolism, means maximum ATP production.

FBA does not require kinetic parameters for each of the chemical reactions, which is ideal for mitochondrial metabolism as these parameters are largely unknown. By altering the constraints on the system, different biological conditions can be modelled. For example, altering both the bounds to be zero on a single reaction prevents any flux through the reaction simulating the inhibition of a single reaction. This feature of FBA makes it an ideal technique for investigating mitochondrial dysfunction, as each complex can be inhibited to identify the potential metabolic adaptions which occur in the system. However, FBA cannot be used to predict precise metabolite concentrations, it can only provide general pathway activity changes. Nonetheless, due to the current lack of understanding of pathway activity during any type of mitochondrial dysfunction, FBA modelling of mitochondrial dysfunction can help provide useful information that, in combination with the results of both previous and future metabolomics studies, would aid in the identification of high confidence biomarkers.

3.1.3. Flux balance analysis models

To accurately predict *in vivo* mitochondrial dysfunction adaptations, an FBA model needs to consist of multiple highly metabolic organs including the brain, liver, kidneys and heart. A multi-organ FBA model of metabolism does not currently exist as most of the regularly maintained models focus on a single organ. However, our understanding of metabolism suggests that organ specific changes in central metabolism are relatively minor. The major functions of each organ, such as the production of urea in the liver, are very well defined and would be easy to implement into the single organ models. The smaller organ specific metabolic differences, such as the differences in amino acid metabolism rates, are less well defined but can be investigated using readily available organ specific gene expression data. Therefore, adapting a single organ model into different organ specific models should be possible. Several different organ models could be generated and subsequently combined into a single, multi-organ model. There are currently two popular FBA

models used to simulate metabolism which are both written in SBML [155] and MIRAM [156] compliant. Both could be used as a base FBA model in the creation of a multi-organ model.

Recon 2.2 is a genome-scale network of human metabolism within a single cell [157]. The model consists of 7,785 reactions involving 5,324 metabolites. The reactions within the model have been separated into nine different cellular compartments including the mitochondrial matrix and mitochondrial intermembrane space. Recon 2.2, and its predecessor Recon 2.0 [158], has been used to successfully model a wide range of different metabolic conditions, such as cancer [159]. The biggest advantage for using Recon 2.2 over other models is the fact that it is a genome-scale reconstruction and hence provides the most comprehensive reconstruction of metabolic adaptations. However, this can also be a large disadvantage as identifying the precise cause and reasoning for a specific metabolic adaptation can be difficult and time consuming. In addition, for mitochondrial metabolism specifically, Recon 2.2 does not consider the electrical gradient component of the proton motive force that is crucial for accurate modelling of OXPHOS activity and ATP production.

MitoCore is a much smaller model focused on modelling the crucial pathways of central metabolism, and by default models cardiomyocyte metabolism [160]. The model consists of 324 reactions and 83 transport reactions with 74 metabolite inputs. The reactions are split into two cellular compartments, the cytosol and the mitochondrial matrix with the intermembrane space considered as cytosolic due to its high permeability. MitoCore has been used to successfully investigate multiple different types of mitochondrial metabolic defects including the deficiency of one of the mitochondrial transporters, the oxodicarboxylate carrier [161], and to investigate the specific role of glutamine during mitochondrial dysfunction [162]. The MitoCore model was specifically created to investigate mitochondrial metabolism so the proton motive force is modelled in its entirety. In addition, the model already contains the reactions of all the major organ specific functions, such as the urea cycle (although this is turned off in the default cardiomyocyte model). The smaller set of reactions makes investigation of the metabolic adaptations much more visible and has proven to provide enough information to enable accurate *in vivo* predictions. The MitoCore

model was therefore used within this study as the base model for creating a multi-organ model for simulating mitochondrial dysfunction.

3.1.4. Chapter summary

In this chapter I describe the process for creating a multi-organ FBA model using a pre-existing single organ model of metabolism, MitoCore. I then use the model to simulate varying levels of inhibition of each of the mitochondrial complexes and discuss the resulting adaptations. The adaptations are evaluated and potential biomarkers for each type of mitochondrial dysfunction, and for varying levels of inhibition, are identified. The chapter starts with simulations of complex III/IV inhibition, since metabolomics data on their inhibition were presented in the previous chapter. The simulations are then extended to modelling inhibition of complexes I and II.

3.2. Methods

3.2.1. Creation of a multi-organ model of human metabolism

The MitoCore model was used as the base for creating four different human organ specific FBA models [160]. These four organs were the heart, which is the MitoCore default model, the liver, the kidney and the brain. These organs were selected as they are the most energetically demanding organs within humans. Major organ specific functions were already written into the MitoCore model, which were turned on for their respective organ models. Minor organ specific alterations in metabolism were investigated using the Human Protein Atlas RNA-Seq organ expression data [30]. Using the organ expression data, reactions were turned on/off in specific organs based on their comparative expression levels across the four organs. For example, the enzyme required for the breakdown of phenylalanine into tyrosine had zero TPM (transcripts per million) in the heart and only 3 in the brain whilst it had a 500 and 720 TPM in the kidney and liver respectively, this reaction was therefore turned off in the brain and heart.

The constraints on the metabolite imports and exports in the MitoCore model are based on experimentally verified results for cardiomyocytes. Experimental data on the constraints of the other organ's does not exist. Therefore, the organ expression data was also used to qualitatively scale these constraints for each organ model, using the cardiomyocyte constraint values as a reference. Each of the organ models was thoroughly checked for errors by forcing a flux through each pathway individually using the objective function of ATP production. The four organ models were then combined into a single model by encasing them in a shared transport system which represented the blood. Similar checks were performed on the complete multi-organ model to ensure the model had no pathway or constraint errors. Model generation was performed using custom Python scripts and the quality checking simulations were performed using the COBRA Toolbox [163].

The default objective function of the MitoCore model was to maximise ATP production which was applicable for all four of the organ specific models. However, in an organism, each of the organs has a different energetic demand with organs such as the brain requiring a larger amount of ATP. The resting energy expenditure rates of each organ has been proposed [164] and confirmed to be accurate in describing the organ specific energy requirements in humans across various age groups [165] for both men and women [166]. The objective function of the multi-organ model was set as ATP production in all four organs with the stoichiometry of each organ's ATP level set based on its resting energy rate. The brain was the most energetically demanding with a stoichiometry of one, the liver the second highest with a stoichiometry of 0.87 followed by the heart at 0.43 and the kidney at 0.4. A complete breakdown of the reactions, their upper bound and lower bound and basal flux value for the multi-organ model can be found in *Supplementary File 3.1*.

3.2.2. Liver mitochondrial complex inhibition

Simulations of liver mitochondrial complex inhibition were performed using the multi-organ model with the COBRA Toolbox and the geometric FBA algorithm to allow for flux profile comparisons [167]. The multi-organ model was simulated under basal conditions to establish the base flux value for all the liver complexes. For each of the

complexes, a simulation was performed at every percent inhibition from zero to one hundred by altering the upper bound of the specific liver complex to lower than its base value. For example, for 10% inhibition of liver complex I its upper bound would be set to 90% of its base flux value. This resulted in 101 different flux profiles for each of the liver complex inhibitions which were comparable. Throughout the process, various upper and lower bounds had to be altered throughout the model to prevent the occurrence of biologically infeasible reactions such as the infinite cycling of metabolite transports to produce co-factors.

3.2.3. MitoCore network visualisation

A network representation of the MitoCore model was generated in Cytoscape using the KEGG database. Each node represented a single compound and each edge represented a reaction with directionally of each reaction indicated by an arrow. The reactions represented by a solid line occurred in the cytosolic compartment of the MitoCore model and reactions represented by a dashed line occurred in the mitochondrial matrix. Each reaction had its fluxes scaled based on the reaction's flux values across all four organs, with a high flux value indicated by red and a low flux by yellow, reactions with zero flux were not assigned an arrow or colour. Compounds were coloured based on generic pathway assignment (*Table 3.1*).

Compound colour	Pathway
Dark blue	TCA cycle
Light blue	Lipid metabolism
Red	Fatty acid metabolism
Orange	Carbohydrate metabolism
Green	Amino acid metabolism
Purple	Purine / Pyrimidine metabolism
Pink	Co-factor biosynthesis
Grey	Other

Table 3.1. The node colour pathway assignments for the network MitoCore representation.

3.3. Results and Discussion

Xenobiotic metabolism is primarily performed in the liver, making it the main location for drug-induced mitochondrial dysfunction. Investigating the organism metabolic adaptations which occur due to liver complex inhibition is therefore critical to furthering our understanding of mitochondrial dysfunction and integral for biomarker identification. The simulated inhibition of each liver complex caused the majority of metabolic adaptations to occur in the liver. Therefore, unless otherwise stated, all pathways and adaptations discussed were based in the liver with the entirety of the model, including the transport system around all four organs which simulates the blood, referred to as the system. The metabolic adaptations are discussed in terms of incrementally increased inhibition from zero percent up to full inhibition. This simulates the accumulation of a drug in the liver and enables the exploration of potential biomarkers for varying levels of inhibition. The complete list of fluxes for all complexes and inhibition levels can be found in *Supplementary File 3.2*.

3.3.1. Modelling liver mitochondrial complex III/IV inhibition

Complex III and IV facilitate the same reaction, they use quinones from the quinone pool to export protons out of the mitochondrial matrix to produce proton motive force, with complex IV ultimately consuming the quinones to produce water from oxygen. These two complexes occur in series during OXPHOS and have no other interactions outside of the OXPHOS pathway. The inhibition of either of these complexes creates the same problem, a huge loss of proton motive force in the mitochondrial matrix which prevents ATP synthase activity, and an inability to remove the free electrons within the inner membrane which are being produced by complex I and II upstream. Therefore, the adaptations which occurred during the flux balance analysis simulations of either complex III or complex IV inhibition were the same. The adaptations which occurred from zero to complete inhibition could be separated into two phases, the fluxes of the pathways identified as important can be seen in *Figure 3.1*. The breakdown of the adaptations into phases was based on the identified major changes in behaviour of the model as the inhibition level gradually increased.

Over the duration of both phases, system ATP synthase levels and the OXPHOS activity of the other three non-inhibited organs began to decrease at one percent and continued to decrease until complete inhibition. In the liver, the same behaviour occurred with complex II, III/IV and ATP synthase. Complex I slightly increased in activity until five percent inhibition then began to decline until complete inhibition. The different behaviour of liver complex I highlights the attempt by the system to rectify the complex III/IV inhibition as complex I was the only unaffected means to produce PMF. However, the quinone by-product prevented the reaction from continually increasing past five percent inhibition of complex III/IV and forced it to decrease in activity like the rest of the OXPHOS complexes. The immediate decrease in system ATP synthase and OXPHOS activity across all the organs highlighted the severity of complex III/IV inhibition as its adaptations were not enough to maintain system ATP levels at any level of inhibition.

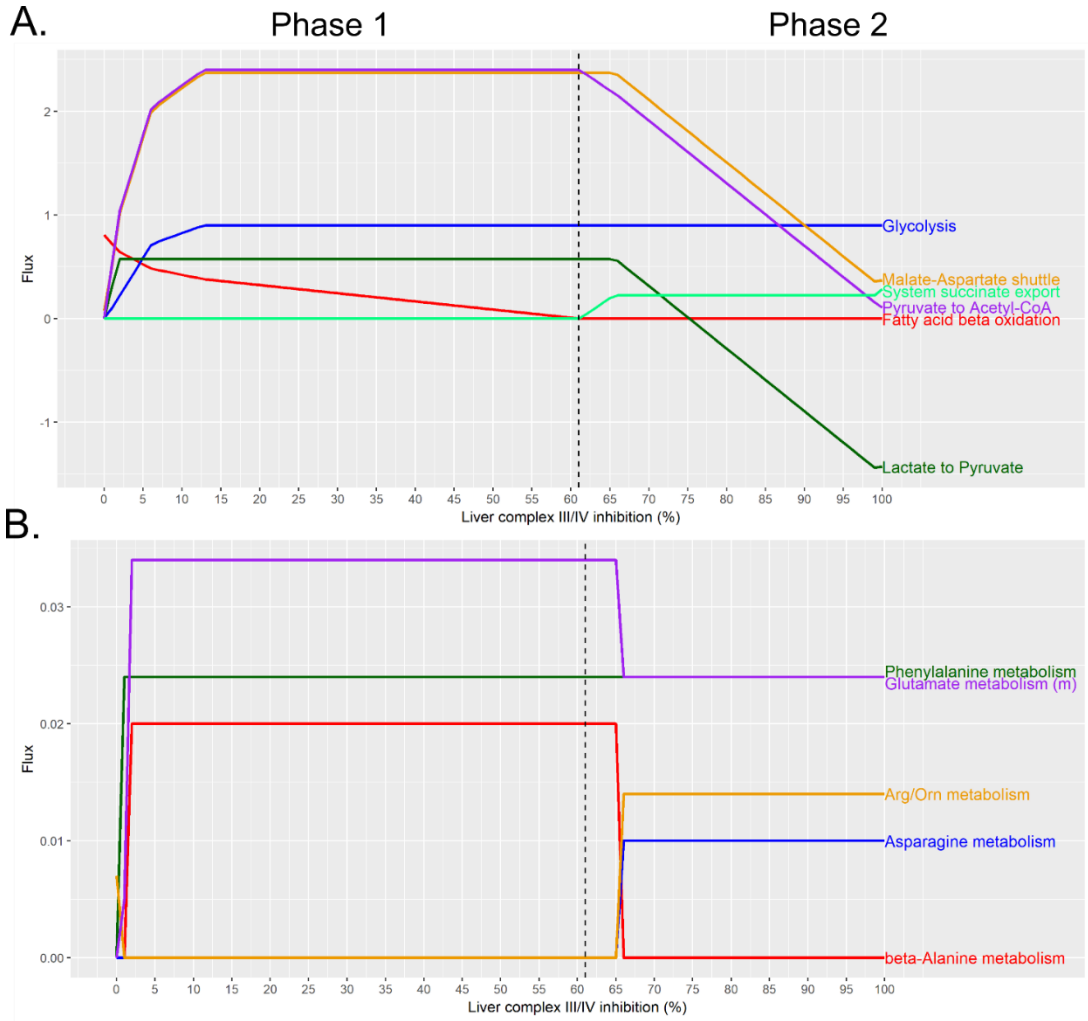


Figure 3.1. Fluxes of important reactions throughout each of the two phases of liver complex III/IV inhibition. A-B are the reactions separated based on their flux magnitude.

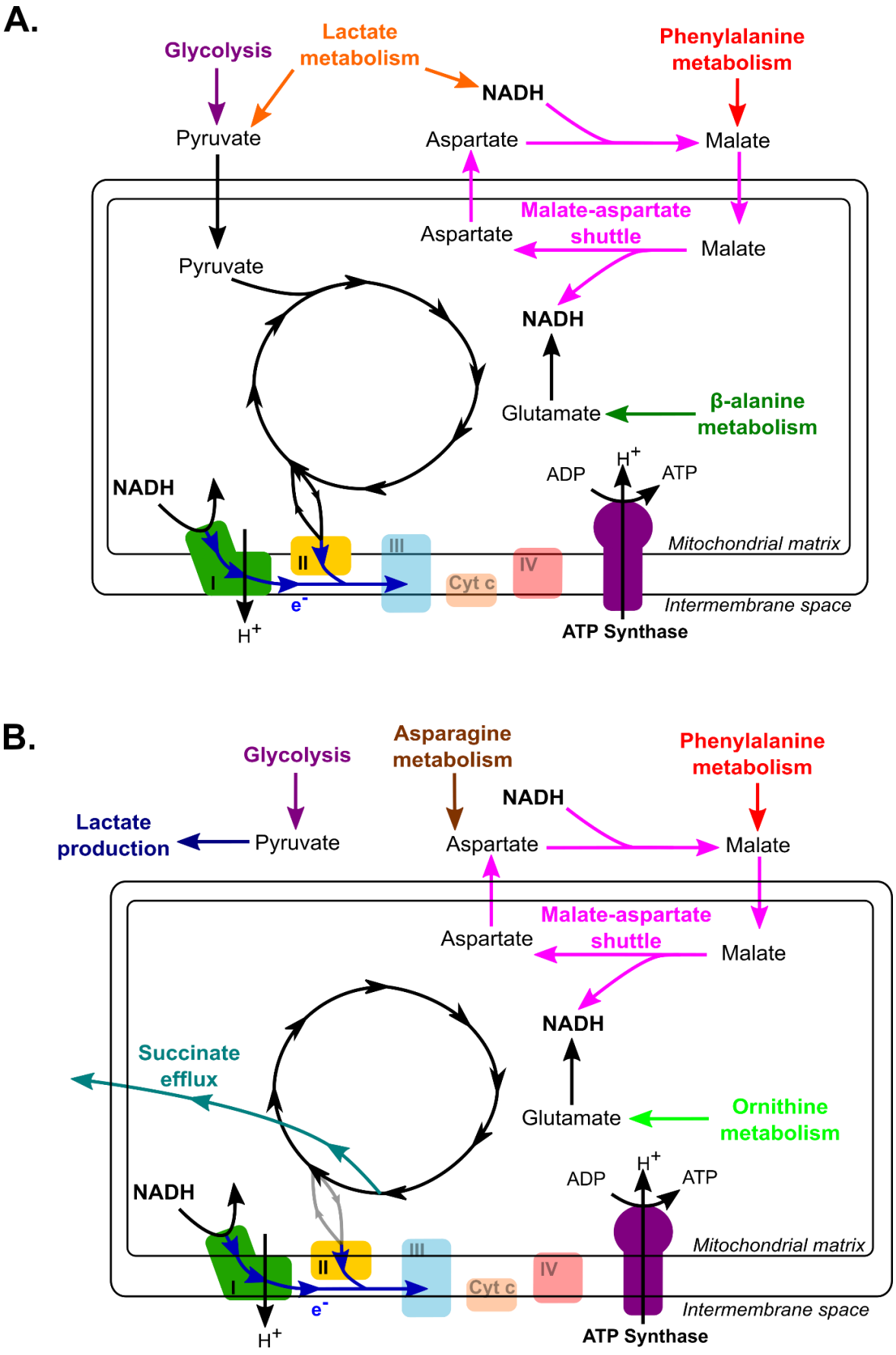


Figure 3.2. Graphical representation of the adaptations which occurred in the liver during each of the phases of liver complex III/IV inhibition. **A)** Phase one. **B)** Phase two.

Phase one

Phase one of the adaptations started at one percent inhibition and continued until sixty-one percent inhibition. ATP production in the liver was immediately supplemented by an increase in glycolysis which continued to increase throughout phase one (*Figure 3.1a, Figure 3.2a*). The pathway reached maximum flux at thirteen percent inhibition which was maintained for the rest of the inhibition levels. The produced pyruvate from glycolysis was shuttled into the mitochondrial matrix, despite costing PMF, and used to feed the TCA cycle via its production into acetyl-CoA (*Figure 3.1a*). This was just one of the adaptations which the system used to increase TCA flux and, by proxy of NADH regeneration, enabled the increased complex I activity to replace the lost PMF from the complex III/IV inhibition. Glucose consumption in the other three non-inhibited organs completely halted during phase one, the heart stopped at twenty percent whilst the brain stopped at thirty-three percent. The kidney began producing glucose via gluconeogenesis at two percent inhibition until twenty percent inhibition where it stopped. Despite being one of the major sites for gluconeogenesis in mammals [168], the only other being the liver, this was the only time that the gluconeogenesis pathway activated in any of the complex inhibition simulations.

The malate-aspartate shuttle transports electrons indirectly across the inner mitochondrial membrane into the mitochondrial matrix (*Figure 3.3*). The shuttle involves two mitochondrial transports; one exchanges 2-oxoglutarate and malate, which was used heavily in the complex II inhibition simulations, and the other exchanges aspartate and glutamate. The net result of the shuttle is the conversion of NADH to NAD+ in the cytosolic compartment of the model and the conversion of NAD+ into NADH in the mitochondrial matrix. The complete malate-aspartate shuttle immediately increased in activity at the beginning of phase one and reached its maximum flux at twelve percent which continued throughout phase one and two (*Figure 3.1a*). The malate to oxaloacetate reaction in the mitochondrial matrix regenerated mitochondrial NADH as part of the TCA cycle. The pathway activity mimicked the behaviour of the pyruvate to acetyl-CoA reaction in the mitochondrial matrix as a collective effort to increase the TCA cycle and complex I fluxes to supplement the lost PMF. The complete activation of the malate-aspartate shuttle

was unique to complex III/IV inhibition, behaviour which could be exploited as a biomarker for complex III/IV inhibition.

The required cytosolic NADH to feed the malate-aspartate shuttle was generated by the conversion of lactate into pyruvate in the cytosolic compartment (*Figure 3.1a*), with the produced pyruvate being transported into the mitochondrial matrix and fed into the TCA cycle. The reaction quickly increased at the start of the phase and reached maximum flux which continued throughout the phase. The increased lactate consumption caused a dramatic increase in the system lactate import, behaviour which would be akin to a decreased level of lactate in a patient's plasma or liver. The metabolomics analysis in the previous section identified a decreased level of lactate in the liver of the rats administered the high dose of the complex III inhibitor. The simulations therefore support this behaviour as a potential biomarker for moderate levels of complex III/IV inhibition. Additional feeding of the malate-aspartate shuttle occurred through phenylalanine metabolism which produced malate in the cytosolic compartment. The pathway immediately activated and reached maximum flux at the start of the phase (*Figure 3.1b*). Intermediates of the phenylalanine/tyrosine metabolism pathway were seen in abundance in the plasma of the rats administered the complex III inhibitor analysed in the previous section, highlighting the potential for these metabolites as biomarkers for complex III/IV inhibition.

Despite being a means for additional ATP production and feeding the TCA cycle along with regenerating mitochondrial NADH, fatty acid β-oxidation immediately decreased in flux throughout phase one (*Figure 3.1a*). The decreased activity was due to the pathway also generating quinones for the quinone pool which are now highly unwanted by the model since their only consumption method is inhibited. This behaviour is entirely unique to complex III/IV inhibition and could be exploited as a biomarker. Fatty acid β-oxidation in the brain slightly increased in flux at the start of the phase but gradually decreased throughout the rest of the inhibition levels. In the heart and kidney, the fatty acid β-oxidation activity continually decreased throughout both phases. The metabolomics analysis in the previous chapter identified a large set of fatty acid β-oxidation intermediates accumulated in both the liver and plasma with many of the intermediates associated with erroneous activity. The undesirable quinone production could be a potential reason for the accumulation of these

metabolites and further highlights the behaviour as a potentially unique biomarker for complex III/IV inhibition.

The attempt by the model to completely remove quinone production throughout the duration of the complex III/IV inhibitions suggests that *in vivo* there would be an abundance of available quinones. Electrons from the quinone pool are known to leak out of the inner mitochondrial membrane into the mitochondrial matrix [169], a process which is important for signalling [170]. An increased leak of electrons has been identified during mitochondrial dysfunction where the increased levels of roaming free radicals are known to cause damage to the mitochondria [122]. More recently, the increasing levels of free radicals during mitochondrial dysfunction have been implicated in altering mitochondrial dynamics during cancer [171]. The analyses of the metabolomics dataset in the previous chapter had clear indications that the rats administered the complex III inhibitor were under an extremely high amount of oxidative stress, and that both cellular and mitochondrial membrane conformational changes were occurring. Therefore, the cause for both behaviours was most likely caused by the increased leak of abundant quinones in the quinone pool caused by complex III/IV inhibition.

Two amino acid metabolism pathways activated immediately in phase one; glutamate and β-alanine (*Figure 3.1b*). Glutamate metabolism regenerated mitochondrial NADH for complex I flux and 2-oxoglutarate which was used to feed the import of malate into the mitochondrial matrix as part of the malate-aspartate shuttle. β-alanine metabolism fed the TCA cycle via acetyl-CoA and produced both mitochondrial NADH and glutamate, which fed glutamate metabolism. During phase one in the brain; leucine, isoleucine and valine all completely switch off. Interestingly, none of these BCAA metabolism pathways activated at all in the liver during complex III/IV inhibition, pathways which were prevalent in the adaptations of both complex I and II inhibitions. In the metabolomics, many erroneous intermediates of the BCAA pathways were identified in the plasma of the rats administered a complex III inhibitor. The BCAA pathways were not activated because of their minor quinone production which was undesirable for the model and could be the potential reason for the presence of these metabolites, highlighting them as potential complex III/IV biomarkers.

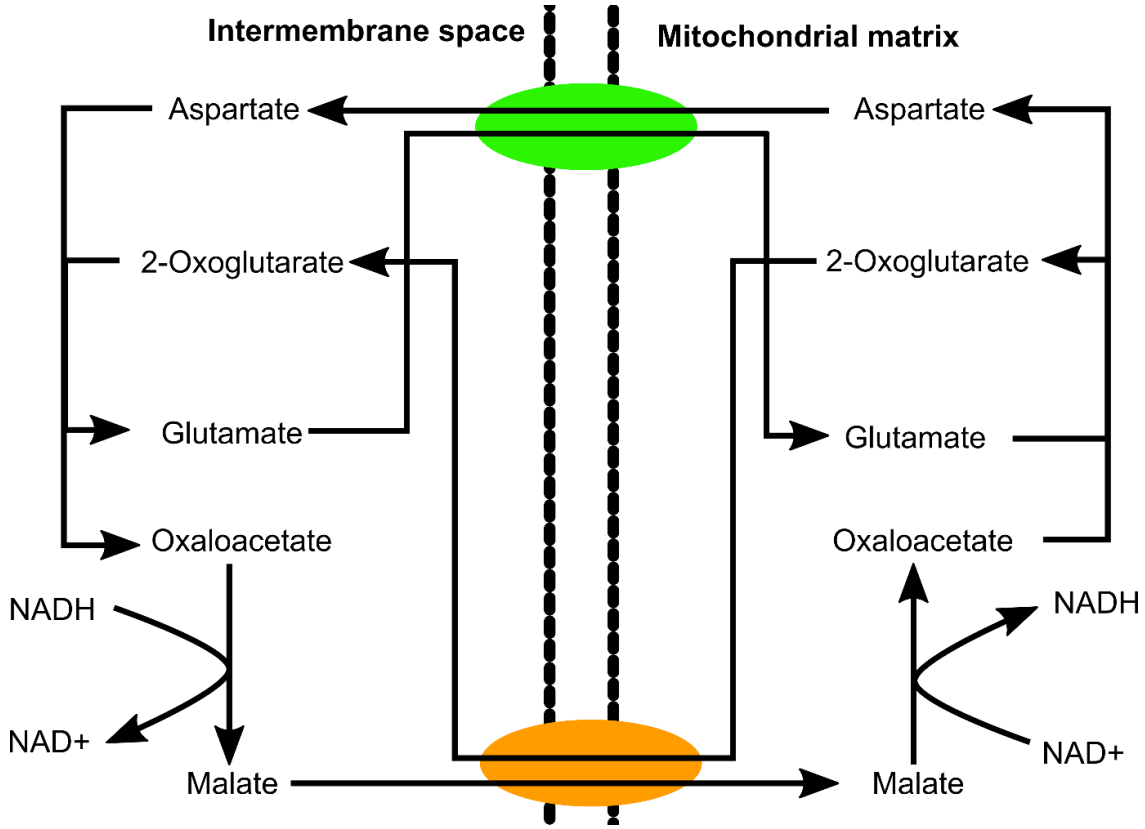


Figure 3.3. The series of reactions which constitute the malate-aspartate shuttle, transporting electrons indirectly across the inner mitochondrial membrane.

Phase two

The gradual blockage of the OXPHOS pathway by the inhibition of complex III/IV led to the complete halt in fatty acid β-oxidation which marked the start of the phase two adaptations (*Figure 3.1a, Figure 3.2b*). As the final reactions of the OXPHOS pathway, the inhibition of complex III/IV also restricted the flux through both complex I and II. The start of phase two marked the tipping point in which the inhibition of complex III/IV overtook the level of mitochondrial NADH as the rate limiting factor for complex I activity necessary for replacing the lost proton motive force. Increased TCA flux and mitochondrial NADH regeneration were therefore in less demand, which caused the progressive decline in flux of the malate-aspartate shuttle and the import and conversion of pyruvate into acetyl-CoA (*Figure 3.1a*). The decreased demand on mitochondrial pyruvate and cytosolic NADH for the malate-aspartate shuttle allowed for the reversal of the lactate to pyruvate reaction which was necessary to remove the now excess pyruvate in the cytosolic compartment

produced by glycolysis. The produced lactate was exported from both the liver and the system. This behaviour is akin to an accumulation of lactate in the liver or plasma of patients, a common feature of known mitochondrial disease which the simulations suggest is a biomarker for only high levels of complex III/IV inhibition.

The reduced flux allowed through complex II prompted the efflux of succinate from both the liver and the system to relieve the pressure put onto the TCA cycle by circumventing the complex II reaction. The efflux occurred immediately into phase two, maximised quickly and remained at its maximum throughout the rest of the phase (*Figure 3.1a*). Similar behaviour occurred during the complex II inhibition simulations, although the efflux was almost nine-fold higher during the complex II inhibitions. At extremely high levels of complex III/IV inhibition (98%) complex II reversed its flux, generating succinate from fumarate which was also effluxed from the system to provide additional relief for the TCA cycle and removing quinones from the quinone pool produced by complex I activity. The accumulation of succinate may therefore be a potential biomarker for high levels of complex III/IV inhibition.

The final predominant metabolic adaptations which occurred at the start of phase two and reached maximum flux instantly were two amino acid metabolism pathways; asparagine and arginine/ornithine (*Figure 3.1b*). Despite declining malate-aspartate shuttle activity, asparagine metabolism was activated to produce aspartate to feed the shuttle. This behaviour could explain the identified decrease in asparagine within the plasma of the rats administered the complex III inhibitor analysed in the previous chapter. The metabolism of β-alanine completely halted at the start of phase two and was replaced by arginine metabolism in the cytosolic compartment which produced ornithine. The ornithine was then transported into the mitochondrial matrix and degraded into glutamate, necessary for continued glutamate metabolism. The swap from β-alanine metabolism to the arginine/ornithine metabolism pathway was caused by the now limited mitochondrial 2-oxoglutarate levels which were instead used in the TCA cycle and effluxed as succinate. This behavioural switch could be used as a biomarker to differentiate between a lower and higher level of complex III/IV inhibition in cases of known complex III/IV inhibition.

After seventy percent inhibition of liver complex III/IV, all four organs remained in phase two and declined in flux until complete inhibition. At complete inhibition, the system ATP production flux decreased by 98% of its basal level.

Rats in the previous chapter were administered a low and high dose of a complex III inhibitor, with the lower dose known to be non-lethal and the higher dose known to be lethal if the animals were left for greater than four hours. In the high dose rats, metabolomics analysis identified decreased lactate in the liver and increased succinate and lactate in the plasma, behaviour which would be expected in phase two of complex III inhibition, placing the inhibition level above sixty percent. Other significant metabolites which suggested the high dose rats were in phase two of the simulation adaptations included increased liver glutamate, ornithine and aspartate (the product of asparagine metabolism). Meanwhile in the lower dose, metabolomics identified no significant difference in any of these metabolites in the liver or plasma but did identify increased liver β-alanine, a metabolite whose metabolism pathway only activated during phase one and suggested that the low dose rats were experiencing less than 60 percent inhibition. The commonalities between the metabolomics and the complex III inhibition simulations provided confidence in the organism model and showcases the way in which pathway switching can be exploited to roughly estimate the level of complex inhibition.

Having modelled inhibition of complex III/IV for which metabolomics data were available, the next simulations modelled inhibition of complex I and II in the liver. These simulations should provide a basis in which *in vivo* studies can be directed to explore mitochondrial dysfunction caused by complex I and II inhibition.

3.3.2. Modelling liver mitochondrial complex I inhibition

Mitochondrial complex I is the first complex of OXPHOS and responsible for oxidising NADH, pumping protons across the inner mitochondrial membrane, and donating electrons to the quinone pool. Its inhibition is expected to affect all these processes. The metabolic adaptations which occurred over the course of complete liver complex I inhibition could be separated into three phases. The reactions which

were identified as important for each of the phases can be seen in *Figure 3.4* and a network representation of the fluxes occurring in each phase was generated (*Figure 3.5*).

Phase one

During phase one (*Figure 3.5a*) the system ATP production, the objective function of the model, did not decrease in flux, the metabolic adaptations which occurred maintained the levels of ATP production across all four organs. The flux through liver complex II-IV slightly increased during the phase, whilst ATP synthase slightly decreased. The increase in the other complexes of the OXPHOS pathway was necessary to generate the proton motive force lost by the complex I inhibition and replace the lost electrons in the quinone pool. However, the decrease in ATP synthase suggested that this adaptation was not enough to fully maintain basal OXPHOS levels. The decrease in ATP production by OXPHOS was met by a slow incremental increase in fatty acid β-oxidation over the duration of phase one (*Figure 3.4c*). The immediate increase in fatty acid β-oxidation as a first response to mitochondrial dysfunction was identified in the metabolomics analysis performed in the previous chapter. This behaviour was prevented in the simulations of complex III/IV inhibition due to its additional function of producing quinones. However, *in vivo* this behaviour is expected in all forms of mitochondrial dysfunction based on our current understanding of metabolism. Therefore, this provided support to the validity of the organism model by replicating the behaviour seen in the metabolomics.

An immediate response to liver complex I inhibition was the rapid decrease in glutamine synthesis and the urea cycle, both of which require ATP and ammonia. This behaviour suggested that these two pathways are the easiest ATP consumers to reduce as an immediate response to complex I inhibition. Glutamine synthesis slowly increased over the duration of phase one after its initial drop in flux (*Figure 3.4a*), most likely necessary to remove ammonia from the liver, whilst the urea cycle steadily decreased, with both pathways completely stopping by the end of phase one. This behaviour adds to the list of potential reasons for the decrease in liver glutamine levels seen in the metabolomics of mitochondrial dysfunction in the

previous chapter. Combined, these adaptations managed to maintain basal liver ATP levels resulting in a steady system ATP production flux over the duration of phase one.

The only change to amino acid metabolism over the duration of phase one was to tyrosine metabolism (*Figure 3.4a*). Tyrosine is metabolised in the cytosol into acetoacetate, a ketone body, and fumarate. The immediate decrease in flux in response to complex I inhibition was most likely due to the by-production of ammonia from the pathway and linked to the decrease in both the urea cycle and glutamine synthesis. The acetoacetate production was necessary for ketogenesis, the ketone body was eventually transported into the brain and metabolised. Ketogenesis in the liver continued at a steady flux across all three phases of complex I inhibition. The fumarate produced was converted into malate and transported into the mitochondrial matrix by the oxoglutarate carrier and used to feed the TCA cycle. Tyrosine metabolism immediately stopped at the end of phase one, behaviour which could be exploited to enable tyrosine intermediates to be used as biomarkers for lower levels of complex I inhibition.

2-Oxoglutarate, also known as α-ketoglutaric acid (α-Kg), is one of the main metabolites involved in the TCA cycle which feeds OXPHOS. During the TCA cycle, 2-oxoglutarate is produced from isocitrate along with either NAD+ or NADP+. The production of 2-oxoglurate in the TCA cycle over the duration of phase one continually swapped from using NAD+ to NADP+ (*Figure 3.4b*). This showed the incrementally reduced amount of available NAD+ during complex I inhibition, as NAD+ is usually produced by the oxidation of NADH at complex I, with the complete lack of availability of NAD+ seeming to trigger the start of phase two. An interesting way the model tried to circumvent this issue involved the cyclic metabolism and transport of alanine and pyruvate.

In the cytosolic compartment, alanine and 2-oxoglutarate were converted into pyruvate and glutamate (*Figure 3.4a*). The pyruvate was then transported into the mitochondrial matrix by the mitochondrial pyruvate transport via proton symport [172], which required mitochondrial proton motive force. This mitochondrial pyruvate, along with glutamate, was then converted into alanine and 2-oxoglutatrate, with the produced alanine transported back into the cytosol, creating a cyclical pathway. The

net product of this pathway was cytosolic glutamate and mitochondrial matrix 2-oxoglutarate, which fed the TCA cycle without the need for co-factors. Therefore, despite only having a small flux, and thus only generating a small amount of 2-oxoglutarate, this identified an interesting interaction with alanine and pyruvate during complex I inhibition. The cyclical pathway activity slowly decreased in flux throughout phase one, most likely due to the increased stress on proton motive force but continued into phase two, where it eventually stopped. The ability to feed the TCA cycle without the need for co-factors would be an ideal therapeutic to deal with complex I inhibition where the NAD+/NADH ratio will be heavily altered. However, the dependence of this pathway on the transportation of pyruvate across the inner membrane means the pathway could never reach high activity, thus invalidating its ability to be supplemented in any way and have any effect on the ratio on a larger scale.

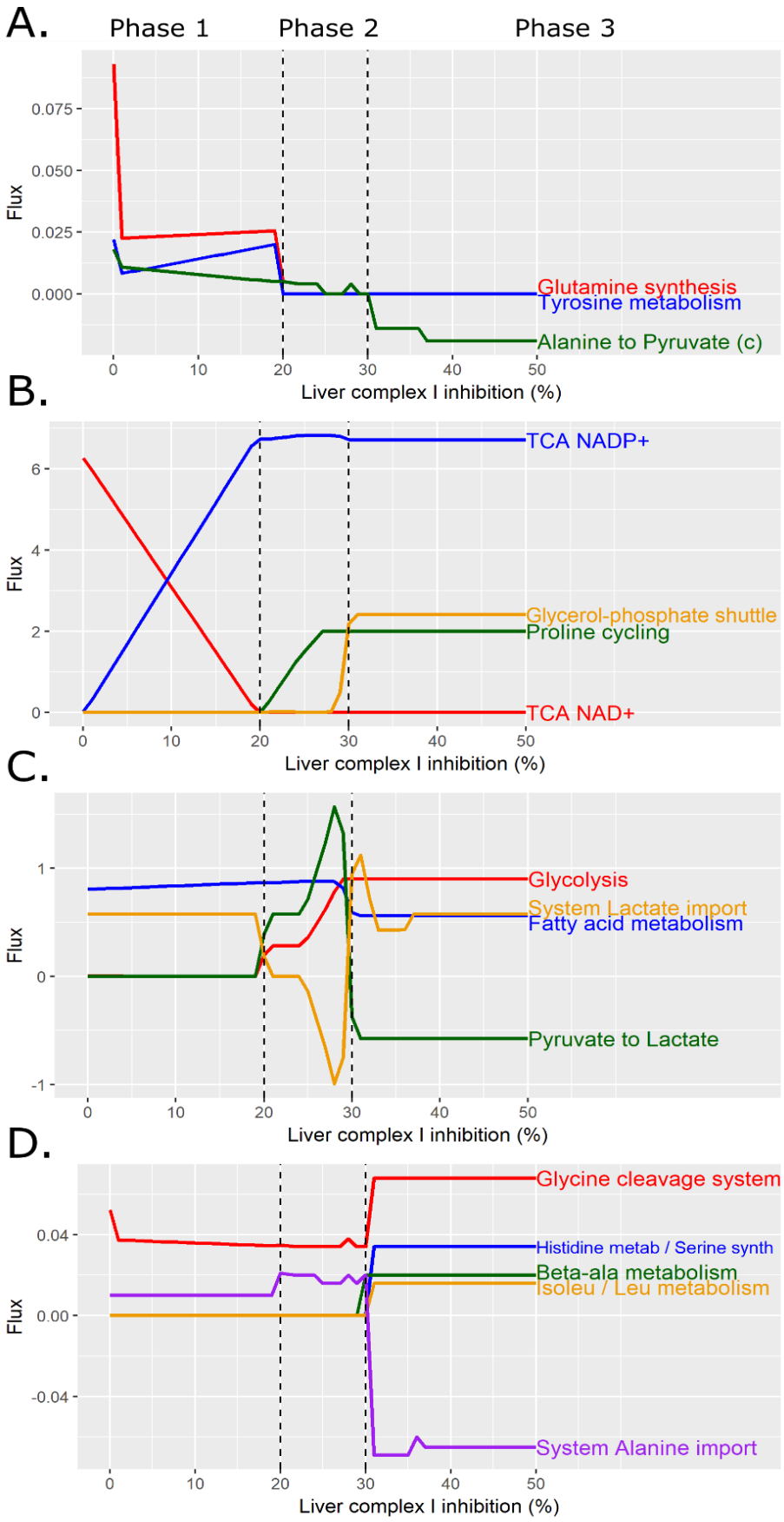


Figure 3.4. Fluxes of important reactions throughout each of the three phases of liver complex I inhibition. A-D are the reactions separated based on their flux magnitude.

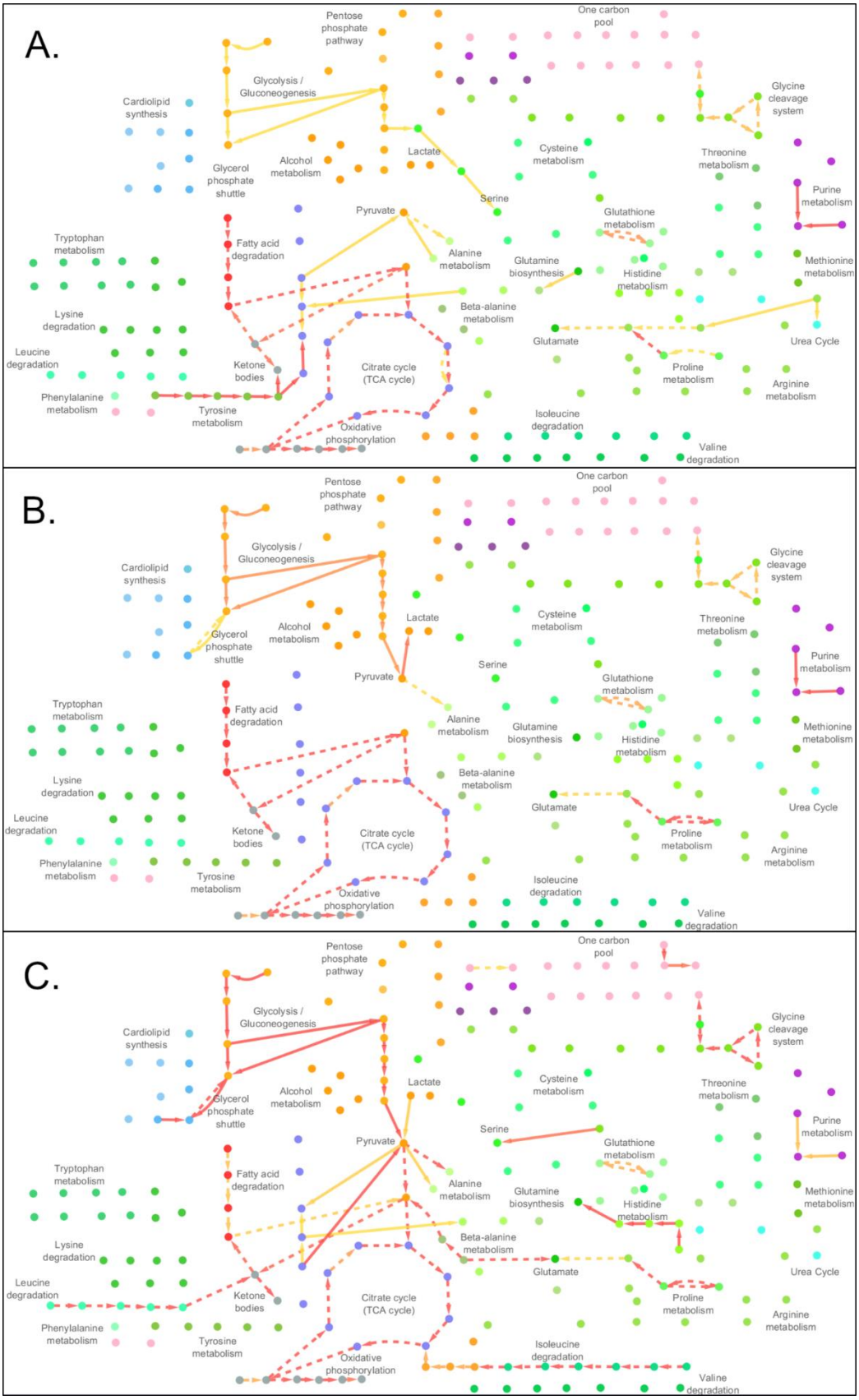


Figure 3.5. Network representations of the flux profiles for each of the three phases of metabolic adaptations for liver complex I inhibition. **A)** Phase one. **B)** Phase two**. C)** Phase three. Edges represent reaction fluxes, with yellow being low and red being a high flux. Nodes represent each of the metabolites, coloured by their generalized pathway (*Table 3.1*).

Phase two

Phase two (*Figure 3.5b*) began at twenty percent inhibition until around thirty percent inhibition with the cause for the switch in behaviour attributed to the heavily perturbed NAD+/NADH ratio. During phase two, liver complex II-IV flux continued to slowly increase and eventually reached their maximum half way through. ATP synthase flux continued to slowly decrease until complex II-IV reached their maximum fluxes, causing the decrease to become much faster. A decrease in ATP synthase activity despite a maintained level of OXPHOS activity suggests that the reduction in ATP production by ATP synthase was caused by the consumption of proton motive force by other reactions, not by its lack of generation by the OXPHOS pathway. Reduced activity of these other proton motive force consumption reactions would allow restored ATP synthase activity and could be a possible therapeutic method. Basal levels of liver ATP production were maintained as fatty acid β-oxidation continued to slowly increase throughout the phase along with the rapid increase in glycolysis, which reached its maximum at the end of the phase (*Figure 3.4c*).

Glycolysis is the conversion of sugars, such as glucose, into pyruvate in the cytosol which requires NAD+ and produces ATP in the process. Under regular conditions, the pyruvate is transported into the mitochondrial matrix by the pyruvate transporter, requiring proton motive force, and used to feed the TCA cycle. However, during mitochondrial dysfunction, proton motive force becomes an expensive resource and the TCA cycle is disrupted. The excess pyruvate produced from the rapid increase in glycolysis during phase two was converted into lactate and exported from the liver, and the system (*Figure 3.4c*). The conversion of pyruvate to lactate also regenerates NAD+ from NADH, giving it the additional benefit of being an NAD+ feeder for glycolysis. The large system export of lactate is akin to the accumulation of lactate in the blood, a hallmark of mitochondrial disease [173]. This result suggests that in the case of complex I inhibition, this accumulation may only occur during a small window of inhibition levels. Behaviour which could be exploited to roughly estimate the level of complex I inhibition in organisms where complex I inhibition is known.

An integral function of complex I activity is the addition of electrons to the quinone pool necessary to enable the proton pumping of complex III and IV for proton motive force generation. Phase two saw the activation of two different pathways which contributed to the quinone pool. Cyclical synthesis and degradation of proline in the mitochondrial matrix began at the start of phase two (*Figure 3.6*) and steadily increased throughout the phase, reaching its maximum flux just over half way through. This pathway requires NADH and produces NAD+, acting as a replacement for complex I by generating mitochondrial NAD+ to feed the necessary TCA cycle reactions, albeit in a much lower capacity. The exact carrier responsible for the transportation of proline into the mitochondrial matrix has yet to be identified but the transportation step is not the limiting factor of proline metabolism in the liver [174]. Supplementation of proline could therefore be explored as a potential therapeutic for complex I inhibition and proline metabolism intermediates used as biomarkers for problems with mitochondrial proton motive force.

Late on in phase two, the glycerol phosphate shuttle rapidly increased in flux. This pathway cyclically converted glycerone phosphate (also known as dihydroxyacetone phosphate (DHAP)) into glycerol-3-phosphate, using NADH as a cofactor and producing NAD+ in the process (*Figure 3.6*). The reaction occurs in the cytosolic compartment and adds quinones to the quinone pool for OXPHOS activity. The increased dependence on NADH for feeding the glycerol-phosphate shuttle forced a decline on the conversion of cytosolic pyruvate into lactate and subsequent decline in the export of lactate from the system. The glycerol-phosphate shuttle is dependent on two enzymes, a cytosolic glycerol-3-phosphate dehydrogenase and a mitochondrial membrane bound FAD-dependant glycerol-3-phosphate dehydrogenase, the activity of the pathway depends on equimolar proportions of both enzymes [175]. Expression levels of these two enzymes are particularly low in liver compared to other organs [176]. However, this does not invalidate the presence of the pathway in the simulations as the primary focus of flux balance analysis is purely qualitative. The rapid activation of the glycerol phosphate shuttle along with the halt in system lactate export marked the end of the phase two adaptations, and the beginning of phase three.

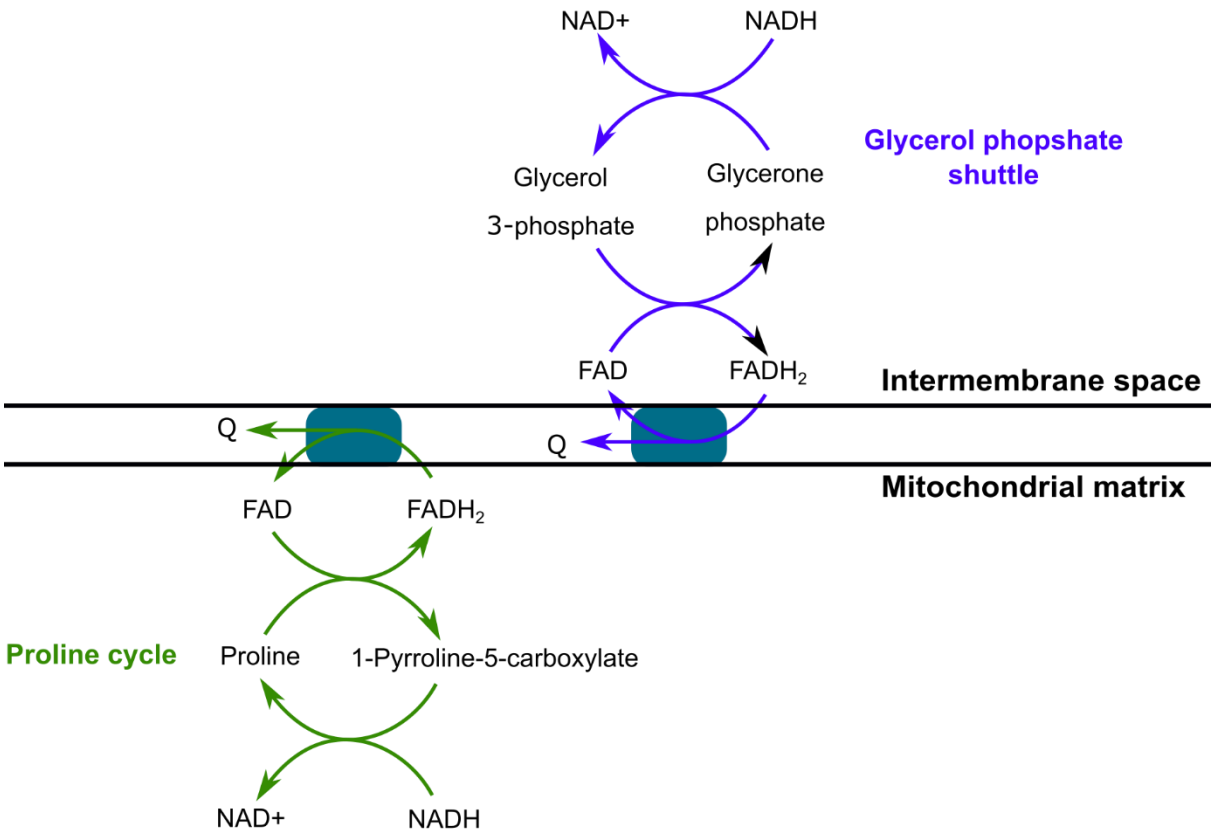


Figure 3.6. Two metabolic pathways that activated during complex I inhibition which contributed to the quinone pool; proline synthesis and degradation in the mitochondrial matrix and the glycerol phosphate shuttle in the intermembrane space.

Phase three

The third phase (*Figure 3.5c*) marked the start of the decline in the system ATP production levels, the metabolic adaptations which occurred could no longer maintain the basal level of ATP production. The metabolic pathways in the three non-inhibited organs began to alter and their complex activity began to decline. The brain, being the highest metabolically demanding organ and largest ATP consumer, underwent the largest changes. The predominate alteration in the brain was the immediate halt of the three branched-chain amino acid (BCAA) metabolism pathways along with threonine metabolism, all of which feed the TCA cycle. In response to the loss of flux feeding the TCA cycle, fatty acid β-oxidation was increased within the brain, although this flux was limited in the model due to the known inefficiency of fatty acid β-oxidation in the brain [177]. Fatty acid β-oxidation within the heart and kidney slowly declined over the duration of phase three, caused by the lowered TCA cycle flux needed to maintain their decreasing complex activity.

The consumption and conversion of glucose into either pyruvate or serine in all three non-inhibited organs had decreased activity over the phase. In the heart and kidney, this meant the immediate halt in all activity whilst the brain had a slow decline in activity until maximum liver complex I inhibition. The adaptation to glucose consumption reflects each organs' metabolic consumption rate. Fatty-acid β-oxidation is the more efficient and slower ATP production method whilst glycolysis is the faster, less efficient pathway. It is therefore expected that, in times of metabolic stress, glycolysis would only be active in the higher metabolically demanding organs or organs under stress, such as the brain and liver in this situation. This result adds credibility to the choice of objective function by accurately reflecting the expected behaviour of glycolysis in human organs put under metabolic stress.

The organism wide metabolic adaptations and the liver supplementation of the quinone pool by the proline cycle and glycerol phosphate shuttle, which both remained at maximum flux, enabled the constant maximisation of liver complex II-IV over the duration of the phase. However, the start of the phase triggered different metabolic adaptations which feed the TCA cycle and subsequent OXPHOS complexes. Fatty acid β-oxidation decreased immediately then remained at the same reduced rate over the phase. To compensate for the loss of TCA feeding, three different amino acid metabolisms were activated; leucine, which fed ketogenesis necessary for constant ketone body degradation in the brain, isoleucine and β-alanine, both of which fed directly into the TCA cycle (*Figure 3.4d*). The third BCAA metabolism pathway, valine metabolism, did not activate, which was the only BCAA metabolite found accumulated in the liver of rats with drug-induced mitochondrial dysfunction studied in the previous chapter. This suggests that during complex I inhibition there was an affinity towards isoleucine metabolism instead of valine metabolism, despite both feeding into the same point of the TCA cycle. The metabolomics study identified metabolites found in each of the three activated amino acid pathways; both leucine and isoleucine intermediates were found in abundance whilst β-alanine accumulated in the liver. All these pathways mark the start of phase three and could therefore be potential biomarkers for high levels of complex I mitochondrial dysfunction. The activation of these specific metabolism pathways could be exploited as a biomarker for high levels of complex I inhibition.

The glycerol phosphate shuttle reaching maximum flux triggered the swap into phase three. Phase three began with the reversal of the cytosolic pyruvate into lactate reaction, necessary to regenerate NAD+ into NADH in the cytosolic compartment to feed the glycerol phosphate shuttle. In accordance, the system lactate export, akin to lactate accumulation in the blood, was stopped and returned to basal import levels. The overall behaviour of the pyruvate to lactate reaction over the course of complex I inhibition makes lactate level monitoring an interesting prospect for complex I inhibition. The complete reversal of the reaction creates an inflection point that should be noticeable in the blood making it an ideal biomarker.

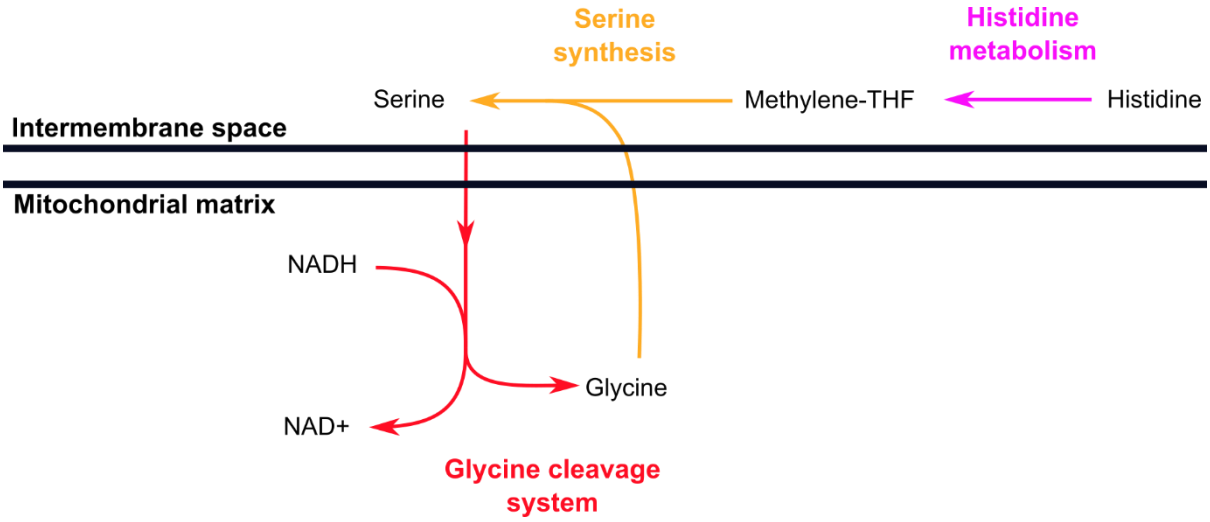


Figure 3.7. The chain of pathways which activated in phase three of liver complex I inhibition, the final adaptations which generated NAD+ in the mitochondrial matrix.

Maximum flux of glycolysis, along with the conversion of lactate into pyruvate in the cytosolic compartment, caused a large availability of pyruvate in the cytosol. For the first time throughout the whole of the complex I inhibition levels, pyruvate was transported into the mitochondrial matrix and converted into acetyl-CoA to replace the lost feeding of the TCA cycle by decreased fatty acid β-oxidation. In both the cytosolic compartment and the mitochondrial matrix, excess pyruvate was converted into alanine and 2-oxoglutarate. The cytosolic 2-oxoglutarate was transported into the matrix by one half of the malate-aspartate shuttle which, along with the mitochondrial 2-oxoglutrate, feed both the TCA cycle and the newly increased BCAA metabolism pathways of isoleucine and leucine. The alanine produced in both compartments was simply exported by the liver and the system (*Figure 3.4d*). This

behaviour, and the cyclical alanine-pyruvate activity which occurred during phase one and two, highlights alanine as a particularly critical metabolite for complex I inhibition. Similar to lactate, the inflection point on the transportation of alanine in the system highlights the monitoring of alanine levels as a potential biomarker for varying levels of complex I inhibition.

The final pathway which activated to regenerate mitochondrial NAD+ was the glycine cleavage system (*Figure 3.7*). The system ultimately converted mitochondrial serine into glycine, generating NAD+ from NADH in the process. The pathway was continually active throughout all three phases but saw an immediate increase in activity as soon as phase three began (*Figure 3.4d*). The increased demand on serine was met by the activation of histidine metabolism which fed directly into serine synthesis via the synthesis of 5,10-methenyltetrahydrofolate (5,10-MTHF). Cytosolic serine was then transported into the mitochondrial matrix and entered the glycine cleavage system. The resulting glycine was then transported out and used to generate more cytosolic serine, creating a cyclical pathway. Ammonia produced during histidine metabolism was transported into the mitochondrial matrix, requiring proton motive force, and used up during the glycine cleavage system, creating a net ammonia neutral reaction at the cost of proton motive force. Intermediates of histidine metabolism were seen in the plasma of the rats suffering from mitochondrial dysfunction analysed in the previous chapter, as was the accumulation of serine in the liver. The behaviour of these pathways' sheds light on potential reasons why these were seen and identified both as potential biomarkers for high inhibition complex I inhibition.

After forty percent inhibition of liver complex I, all four organs remained in phase three and declined in flux until complete inhibition. At complete inhibition, the system ATP production flux decreased by 26% of its basal level.

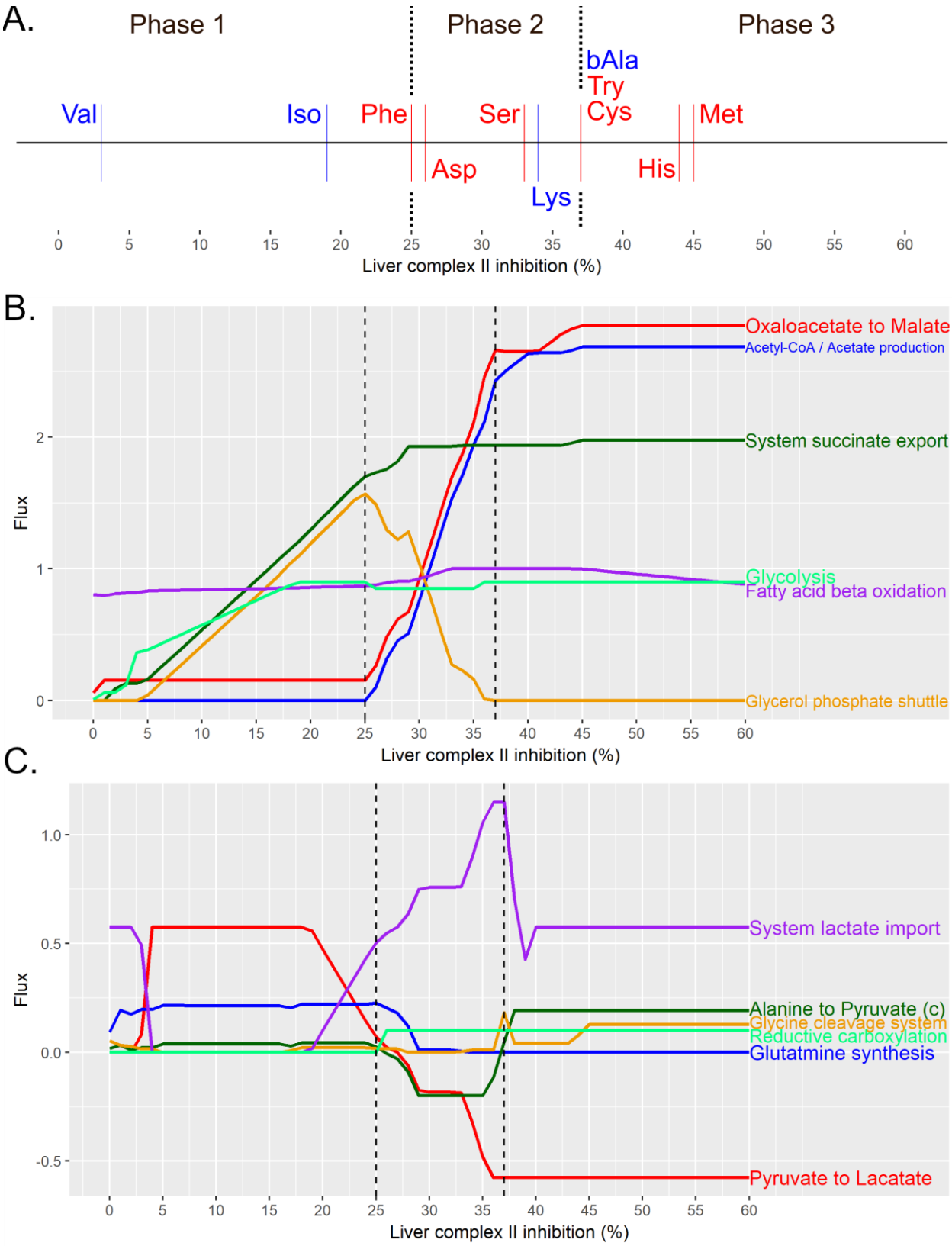


Figure 3.8. A) The activation points of each amino acid metabolism pathway from zero to complete inhibition of liver complex II which maximised in flux upon activation. Blue for pathways which occur in the mitochondrial matrix, red for those in the cytosolic compartment. **B/C)** Fluxes of important reactions throughout each of the three phases of liver complex II inhibition.

3.3.3. Modelling liver mitochondrial complex II inhibition

Mitochondrial complex II is part of both the OXPHOS and TCA pathways, and its inhibition is expected to affect both. Multiple amino acid metabolism pathways can be used to feed the TCA cycle such as glutaminolysis, the conversion of glutamine into 2-oxoglutarate. Many of the adaptations which occurred during complex II inhibition were the activation of various amino acid metabolism pathways, their point of activation over the course of the complex I inhibition simulations are shown in *Figure 3.8a*. Once activated, all these pathways immediately reached maximum flux and maintained their flux throughout the rest of the complex II inhibition simulations. The fluxes of other reactions which were identified as important metabolic adaptations during complex II inhibition are shown in *Figure 3.8* and were separated into three different phases.

Phase one

Phase one began at one percent inhibition and ran until around twenty five percent inhibition. During this phase, the system ATP production level was maintained, suggesting adequate adaptations to cope with the increased stress. The activity of the liver OXPHOS pathway maintained its basal level with ATP production immediately supplemented by glycolysis which slowly increased over the phase and eventually reached its maximum (*Figure 3.8b*). The spike in cytosolic pyruvate levels from glycolysis was paired with its conversion into lactate, preferred over its import into the mitochondrial matrix which costs proton motive force (*Figure 3.8c)*. Similar behaviour was seen during the complex I inhibition simulations once glycolysis began. However, unlike the complex I inhibition simulations, the lactate did not get exported out of the system. Instead, the lactate was used to feed the other organs, which convert it back into pyruvate to feed their own TCA cycles. Therefore, this pathway activity would not create an accumulation of lactate in the plasma of organisms with low levels of complex II inhibition but may cause increased levels of lactate in the liver.

One of the important functions of a complete TCA cycle is the regeneration of mitochondrial NADH from NAD+ that is necessary for complex I activity. As one of the reactions of the TCA cycle, complex II inhibition caused a blockage in the cycle making it necessary to find alternate ways to regenerate NADH. Valine metabolism was the first reaction activated which fulfilled this role (*Figure 3.8a*). Additionally, valine metabolism generates a small amount of quinones for the quinone pool and feeds the TCA cycle (*Figure 3.8*). At three percent inhibition, the pathway switched on and immediately reached maximum flux. This showed a contrast to the complex I inhibition simulations where valine metabolism never activated, highlighting the different amino acid metabolism affinities of the system to different complex inhibitions. Valine metabolism intermediates could therefore be used to differentiate between mitochondrial dysfunction caused by complex II inhibition.

The supplementation of the TCA cycle by valine metabolism and the blockage caused by the inhibition of complex II necessitated the efflux of succinate, an intermediate of the TCA cycle and complex II reaction (*Figure 3.8b*). Succinate was exported from both the mitochondrial matrix and system, akin to an accumulation of succinate in the blood (*Figure 3.9*). Over the duration of all three phases of complex II inhibition, succinate export from the system continued with a steadily increased efflux during phase one and two, where it reached its maximum. The accumulation of succinate has been seen *in vitro* when cells are given known complex II inhibitors [178,179]. In addition, an increased level of succinate has been proposed as a diagnostic marker for cancers which have altered complex II activity [180]. Therefore, an increase in succinate levels could be used as a biomarker for complex II inhibition with the fold-change increase being representative of the inhibition level at lower levels of inhibition.

The loss of complex II activity caused a decrease in quinone production for the quinone pool. In addition to valine metabolism activation, the glycerol phosphate shuttle was also activated to supplement the quinone pool at the same complex II inhibition level (*Figure 3.8b*). The pathway steadily increased in flux throughout phase one and reached its maximum flux at the end of the phase. The pathway activated much earlier than during complex I inhibition and increased much slower. This highlights a different affinity for the major quinone generation pathways in

complex II inhibition when compared to complex I, caused by the difference in secondary issues related to their inhibitions. Early detection of glycerol phosphate shuttle intermediates could therefore be used to differentiate complex II inhibition.

Further into the phase at around seventeen percent inhibition, isoleucine metabolism activated in the mitochondrial matrix (*Figure 3.8a*). The pathway fulfilled a similar function to valine metabolism by feeding the TCA cycle, regenerating mitochondrial NADH and contributing a small amount of quinones to the quinone pool (*Figure 3.9*). An increase in valine and isoleucine metabolism put an increased demand on mitochondrial 2-oxoglutarate, required for the first steps in both pathways. Throughout the complex II inhibition phases, mitochondrial 2-oxoglutarate levels were supplemented by the conversion of pyruvate into alanine and 2-oxoglutratate, with the excess alanine produced transported back into the cytosolic compartment. This put increased demand on the transportation of cytosolic pyruvate into the mitochondrial matrix. Combined with the growing cytosolic NADH demands of the glycerol phosphate shuttle when isoleucine metabolism activated, and the increased demand for pyruvate to feed the TCA cycle, caused a slow decrease in lactate production from cytosolic pyruvate. Interestingly, the feeding of the TCA cycle by pyruvate in the mitochondrial matrix was done exclusively by its conversion into oxaloacetate during phase one instead of its regular conversion into acetyl-CoA, despite this reaction costing ATP. This was because the levels of mitochondrial oxaloacetate became the limiting factor in TCA cycle flux due to the blockage caused by complex II.

Phase two

Phase two began at twenty five percent inhibition and lasted until around thirty seven percent inhibition of complex II. The adaptations during this phase still managed to cope with the added stress, maintaining the basal level of system ATP production. The complexes in the liver, except for complex II, slowly increased in flux during the phase and reached their maximum flux just before the swap into the third phase. With glycolysis reaching its maximum flux, the adaptations during phase two focused on increasing OXPHOS activity and relieving the blockage in the TCA cycle.

Additional ATP was produced by the steady increase in fatty acid β-oxidation which maxed out just over half way through the phase (*Figure 3.8b*). The blocked TCA cycle caused the system to prefer maximising glycolysis first over fatty acid β-oxidation as the pathway occurred in the cytosolic compartment and could produce ATP without interacting with the TCA cycle. The opposite behaviour occurred during complex I inhibition, where fatty acid β-oxidation was the first pathway to reach maximum flux, behaviour which could be used to differentiate complex II inhibition *in vivo*.

The trigger for swapping into phase two was the activation of the production of cytosolic malate from oxaloacetate which required cytosolic NADH (*Figure 3.10*). The steady increased demand on cytosolic NADH caused the decline of the glycerol phosphate shuttle over the duration of phase two which completely stopped by the end of the phase (*Figure 3.8b*). The produced cytosolic malate was transported into the mitochondrial matrix by the mitochondrial citrate carrier, which exchanged mitochondrial citrate for cytosolic malate (*Figure 3.9*) [181]. The mitochondrial malate then entered the TCA cycle where it regenerated mitochondrial NADH. The eventual product of the imported malate was mitochondrial citrate as part of the natural progression of the TCA cycle which was used to import more cytosolic malate via the citrate carrier. Combined with the increased efflux of succinate from the mitochondrial matrix, the behaviour allowed greater flux in the TCA cycle by acting as a bypass around the inhibited complex II reaction of the TCA cycle. The increased TCA flux regenerated more mitochondrial NADH resulting in the slow increase in OXPHOS activity seen throughout the phase. Cytosolic malate levels were immediately supplemented by the activation of phenylalanine metabolism and asparagine metabolism in the cytosolic compartment (*Figure 3.8a, Figure 3.10*). Intermediates of both these pathways were found decreased in the plasma of the rats studied in the metabolomics data set.

The exported citrate was then combined with CoA in the cytosolic compartment to generate acetyl-CoA and oxaloacetate, closing the loop to create a cyclical pathway for feeding the TCA cycle (*Figure 3.10*). The cytosolic acetyl-CoA then degraded into CoA, creating a net neutral use of cytosolic CoA, and acetate which was exported from the liver and system. The flux through all the pathways involved in this cycle

rapidly increased throughout phase two and continued to increase in phase three. The rapid efflux of acetate in this manner would cause the accumulation of acetate in patients with complex II dysfunction, although this has not been identified in any previous studies. An alternative clinical phenotype for this cyclical pathway would be the accumulation of either acetyl-CoA in the liver or citrate in the blood with the production of acetate either not active or rate limited. Nevertheless, this behaviour highlighted three different metabolites which could be used as biomarkers for identifying the circumvention of the TCA cycle using the citrate carrier, caused by a blockage in the TCA cycle or complex II inhibition. The rats studied in the previous chapter had an increased level of citrate in their blood.

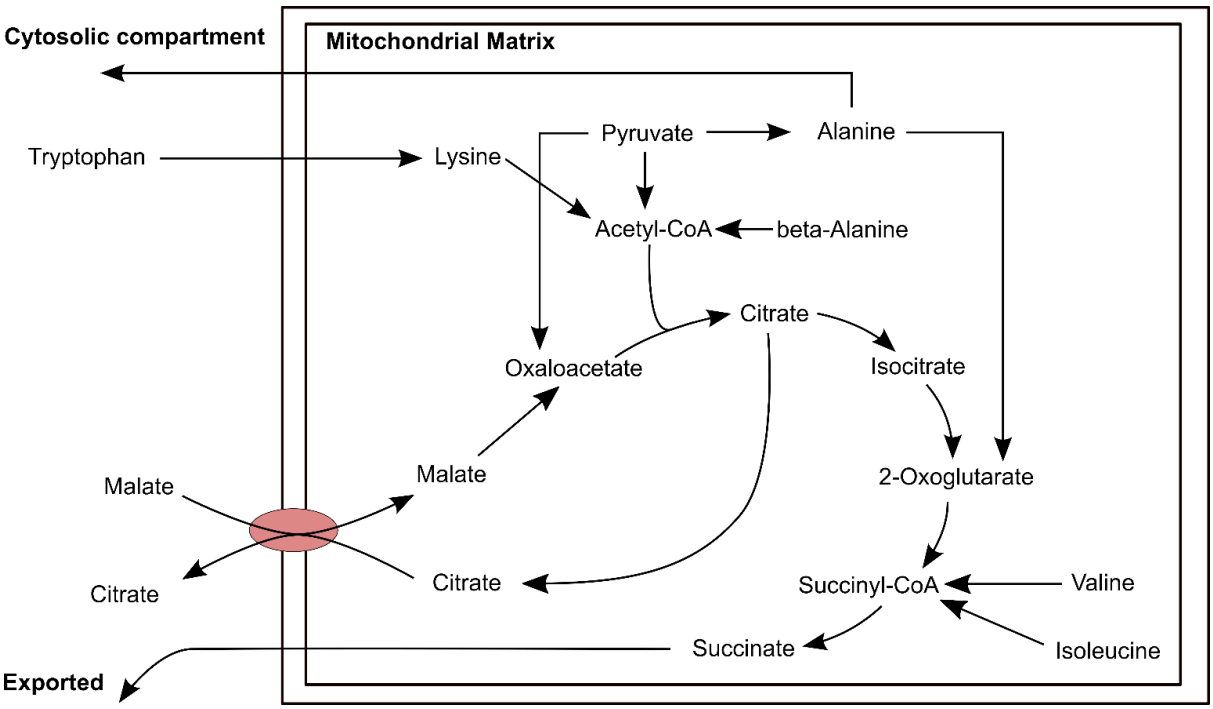


Figure 3.9. The metabolic adaptations in the mitochondrial matrix which fed the TCA cycle over the course of zero to complete inhibition of liver complex II.

Additional cytosolic citrate, used in the production of oxaloacetate, was generated using reductive carboxylation which immediately activated and reached its maximum flux once phase two started (*Figure 3.8c*), converting cytosolic 2-oxoglutarate into citrate. The increased demand on 2-oxoglutarate was fulfilled by the conversion of cytosolic pyruvate into alanine and 2-oxoglutarate (*Figure 3.10*), which continued throughout the phase (*Figure 3.8c*). This was a reversal of the basal directionality of the pathway which throughout phase one was converting alanine into pyruvate.

Paired with this reaction in phase one was the synthesis of glutamine which was being used to remove excess cytosolic glutamate produced as a by-product of the alanine to pyruvate conversion (*Figure 3.8c)*. The synthesis of glutamine therefore completely stopped during phase two after the steady increase in alanine production. The alanine produced in the cytosolic compartment of the liver during phase two was exported from both the liver and the system. Clinically this would present itself as an accumulation of alanine in the plasma during complex II inhibition, albeit only for the small window of inhibition represented by phase two.

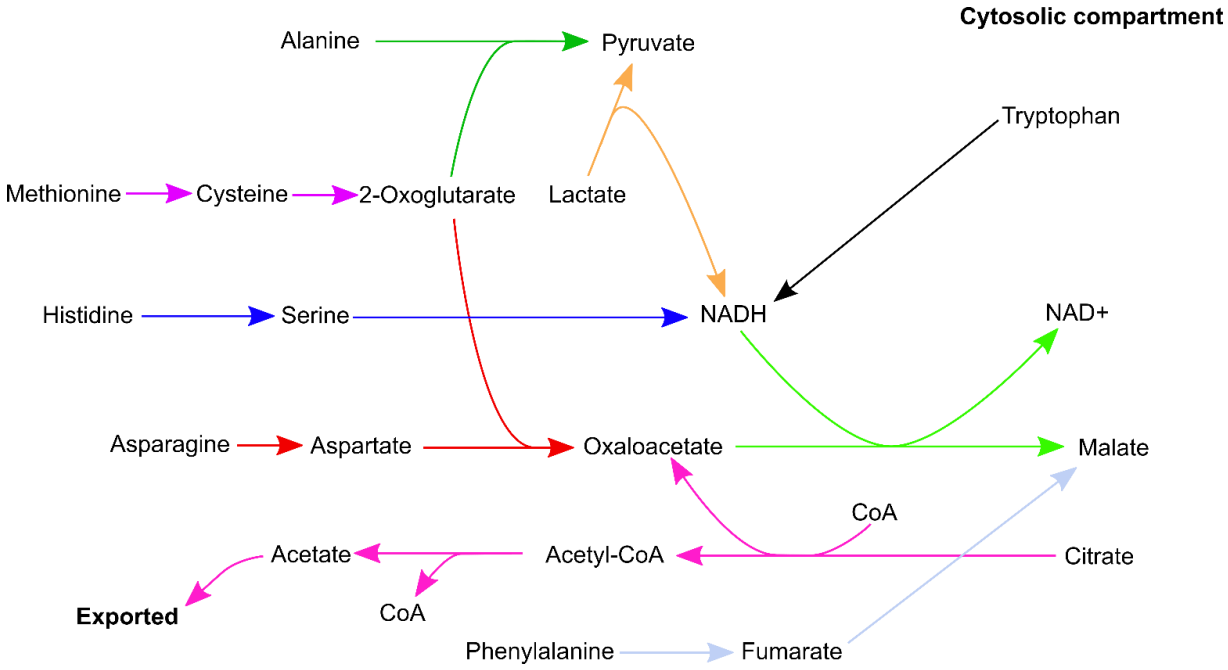


Figure 3.10. The metabolic adaptations in the FBA model cytosolic compartment which fed the production of malate over the course of zero to complete inhibition of liver complex II to facilitate the malate-citrate exchange feeding the TCA cycle.

Further cytosolic NADH regeneration to feed malate production began at the start of phase two using the conversion of lactate into pyruvate (*Figure 3.10*). The flux steadily increased during the phase and eventually reached its maximum towards the end of the phase (*Figure 3.8c*). The constant consumption of lactate caused an increase in the system import of lactate which would be equivalent to a lactate deficiency in the plasma. A decrease in liver lactate level was identified in the rats with known mitochondrial dysfunction studied in the previous chapter. The overall

behaviour of the pyruvate-lactate pathway over the course of the complex II inhibition simulations identified lactate as a potentially useful biomarker. The rapid increase in production, subsequently followed by a steady increase in the reverse reaction would be expected to cause a detectable change in plasma levels over the course of varying complex II inhibition levels, allowing for a rough estimate of an organism's level of complex II inhibition.

The final two metabolic adaptations which occurred during phase two were the activation and immediate maximisation of serine biosynthesis and lysine metabolism. Serine biosynthesis in the cytosolic compartment was used to generate cytosolic NADH necessary to fuel malate production (*Figure 3.10*). Lysine metabolism in the mitochondrial matrix fulfilled a similar role to the two BCAA metabolism pathways already active in the mitochondrial matrix, it regenerated mitochondrial NADH for complex I activity, fed the TCA cycle via acetyl-CoA and generated a small amount of quinones which were added to the quinone pool (*Figure 3.9*).

Phase three

The switch into phase three began at around thirty seven percent inhibition, triggered by the return to the basal directionality of the alanine to pyruvate reaction in the cytosolic compartment due to increased demand on pyruvate to feed the TCA cycle. System ATP production levels began to decline for the first time in the complex II inhibition simulations along with a decrease in complex activity of the non-inhibited organs. In the heart and kidneys, all glucose consumption completely stopped whilst in the brain, leucine, isoleucine, valine and threonine metabolism stopped which was met by a slight increase in fatty acid β-oxidation. All these adaptations were the same adaptations seen in the complex I inhibitions, just at a slightly higher inhibition level. Liver OXPHOS activity slowly decreased over the duration of the phase. Liver ATP synthase activity rapidly decreased at a higher rate than the other complexes due to a necessary increased import of pyruvate into the mitochondrial matrix to feed the TCA cycle which, for the first time, was via its conversion into acetyl-CoA.

The activity of the alanine to pyruvate reaction in the cytosolic compartment during phase three was much higher than its basal level which occurred during phase one (*Figure 3.8c*). This activity was matched by an increased system and liver import of alanine which would most likely be seen clinically as a change in alanine levels in both the liver and blood. The other metabolite for this reaction was cytosolic 2-oxoglutarate, putting a much higher demand on the metabolite. In response, the system immediately activated and maximised cysteine metabolism (*Figure 3.8a*). Further into the phase, and the last adaptation which occurred in the liver for complex II inhibition, was the activation of methionine metabolism which supplemented cysteine metabolism and cytosolic 2-oxoglutarate production necessary for both alanine and asparagine/aspartate metabolism (*Figure 3.10*).

The other three metabolic adaptations which occurred during this phase were all amino acid metabolism pathways. In the cytosolic compartment, tryptophan metabolism was activated immediately into the phase which was used to regenerate cytosolic NADH for the continued activity of the malate-citrate exchange (*Figure 3.8a*), the cyclical pathway seen in phase two which continued its activity throughout the whole of phase three (*Figure 3.10*). The final steps of tryptophan metabolism occurred in the mitochondrial matrix which shares the same final reactions as lysine metabolism, both of which fed the TCA cycle and produced a small amount of quinones for the quinone pool. At the same inhibition level, β-alanine metabolism was activated in the mitochondrial matrix which regenerated mitochondrial NADH and fed the TCA cycle via acetyl-CoA (*Figure 3.9*). The penultimate pathway to activate was histidine metabolism in the cytosolic compartment, paired with an increase in the glycine cleavage system to ensure the pathway was ammonia neutral (*Figure 3.8c*). Histidine metabolism was used to ultimately produce cytosolic serine for cytosolic NADH regeneration (*Figure 3.10*).

A plethora of the amino acid pathways which activated throughout the complex II inhibition simulations were not present in the complex I and III/IV simulations. In fact, complex I had its own unique amino acid pathway activated, leucine metabolism. *Figure 3.8a* represents the ordered affinity of the system for each of the different amino acid metabolisms which switch on during progressively higher complex II inhibition in the liver. Therefore, the monitoring of intermediates within each of these

pathways could act as multiple biomarkers for identifying the level of complex II inhibition present in an organism. For example, identifying an increased activity in both valine and asparagine metabolism, but not tryptophan metabolism within an organism would indicate an inhibition level of less than thirty seven percent and place them in phase two of adaptations.

After fifty percent inhibition of liver complex II, all four organs remained in phase three and declined in flux until complete inhibition. At complete inhibition, the system ATP production flux decreased by 61% of its basal level.

3.4. Conclusion

The simulations of all the complex inhibitions in the liver highlighted an obvious difference in the ability of the system to adapt to each complex inhibition and their subsequent lethality. Complete inhibition of complex I reduced the basal level of system ATP production by a quarter, whilst complex III/IV inhibition reduced the production to almost zero. Being the first step in the pathway and having no other interactors, complex I inhibition had multiple adaptations which counteracted the lost function and managed to maintain basal levels of ATP production until twenty five percent inhibition. Complex II inhibition caused a larger number of adaptations than complex I activity which primarily aimed to fix the blockage in the TCA cycle caused by lost complex II activity. The adaptations maintained the basal ATP production level until thirty seven percent inhibition but ultimately resulted in a sixty percent decrease in system ATP production once completely inhibited. Complex III/IV inhibition, being the last stage in the OXPHOS pathway and having no other interactors, had the smallest number of adaptations and caused the most drastic decrease in system ATP production levels at complete inhibition. The system had no means to avoid the complex III/IV inhibition and the reduced ability to safely remove quinones from the quinone pool restricted the possible adaptations causing the system ATP production to begin reducing as early as one percent inhibition. Complex III/IV inhibition is markedly more lethal to a system than complex I & II inhibition.

The unique behaviour of the two primary ATP production replacement pathways, glycolysis and fatty acid β-oxidation, over the duration of each complex inhibition highlights a simple mechanism for identifying which complex is inhibited in cases of known mitochondrial dysfunction during drug development. Complex I inhibition saw an increase in fatty acid β-oxidation as a first response, followed by an increase in glycolysis paired with a drop in fatty acid β-oxidation. Complex II inhibition saw the rapid increase in glycolysis as a first response followed by a slow increase in fatty acid β-oxidation. Finally, complex III/IV saw a rapid increase in glycolysis as a first response with a decline and complete halt of fatty acid β-oxidation. Therefore, the monitoring of these two pathways over various drug concentration levels would enable the identification of the complex being inhibited by the drug, a theory which should be explored in future experimentation.

The simulations also identified many other pathways and behaviours which were unique to each complex inhibition that could be used to differentiate between different types of mitochondrial dysfunction caused by complex inhibition. Complex I inhibition can be identified by its drastically increased activity of proline cycling and the glycerol phosphate shuttle to replace the lost quinone production, along with the increase in a small set of amino acids at relatively high levels of inhibition. Complex II had a unique efflux of succinate and increased malate-citrate exchange creating an extended TCA cycle to avoid the complex II reaction. In addition, complex II inhibition had a plethora of amino acid metabolism pathways which activated at different inhibition levels, giving the ability to estimate a complex II inhibition level based on which pathways were switched on or off. Finally, complex III/IV had a unique increase in the complete malate-aspartate shuttle along with a drastic avoidance in quinone production which the metabolomics suggests could be causing the accumulation of biomarkers which indicate erroneous activity of pathways that generate quinones as a by-product, such as BCAA metabolism and fatty acid β-oxidation.

Other particularly interesting results of the simulations point to two specific metabolites which had particularly dynamic behaviour; alanine and lactate. Alanine is closely related to pyruvate and had highly dynamic behaviour in both complex I and II inhibition where it swapped from being consumed to produced, behaviour which

would be ideal in estimating a precise complex inhibition level. Lactate accumulation is a known symptom of mitochondrial dysfunction however, these simulations suggest that the case may not be as simple as believed. The metabolite, which also has a close relationship with pyruvate, swapped behaviour from consumption to production during all three complex inhibition simulations, highlighting it as an interesting prospect as a biomarker for identifying both the type of complex inhibition and an estimated inhibition level.

The organism model used for the simulations used constraints based on qualitatively adapted values taken from measurements in cardiomyocytes. Although these are expected to be within reasonable bound of the true values for each of the different organs, they can be massively improved upon. As a result, the discussed adaptations are of a qualitative nature, the different phases of adaptations may occur at different inhibition levels *in vivo* and may occur during a different inhibition window than identified in the simulations. Future studies focused on measuring each organ import/export rate of the various metabolites used in the model would improve simulation accuracy and allow for greater confidence in precise inhibition levels and phase duration windows. Further simulations of different types of mitochondrial dysfunction using the organism model, such as inhibiting complexes in a different organ or inhibiting multiple complexes at once, should also be performed in the future. These simulations would build on the results of this study by creating a larger basis in which unique and shared behaviours can be identified, allowing for the identification of biomarker which could confidently differentiate between types of mitochondrial dysfunction and give an estimate of their inhibition level.

Chapter 4

Predicting mitochondrial localisation

4.1. Introduction

4.1.1. Mitochondrial protein localization

Advancements in next generation sequencing over the last decade has led to the increased knowledge of the human genome with almost 21,000 gold-standard reviewed human putative protein-coding genes listed in the UniProt database [182]. Less than nine hundred of these are confidently localised in mitochondria with an additional six hundred predicted, but currently unknown [28]. The critical functions of the mitochondria and their involvement in serious human diseases makes the set of unknown proteins an important piece of knowledge. The set of proteins could contain the missing link in allowing us to generate therapeutics for multiple diseases such as Alzheimer's disease. Many experimental attempts have been made to complete the mitochondrial proteome.

The most common experimental technique involves using mass spectrometry, and many large-scale studies have been done using multiple tissues in a range of species [183–185]. The process involves fractioning the mitochondria from cells and putting the protein contents of the mitochondria through a mass spectrometer using the latest technology, such as liquid chromatography or tandem mass spectrometry. The mass spectrometer identifies which proteins are present in the mitochondria resulting in binary data sets where each protein is labelled as either present or absent. The main problem with the technique is contamination resulting from either human error or difficulties in fractioning the mitochondria from the cell, leading to

proteins being incorrectly annotated as mitochondrially localised, e.g. keratin which is commonly transferred from the environment [186].

Alternative experimental techniques for identifying mitochondrial proteins involve imaging complete cells and tracking protein localisation. The experiments are performed by tagging a protein of interest with another protein that has fluorescent properties making it visible under a confocal microscope, which allows for the tracking of protein localisation in the cell. The tagging is done by either anti-body staining [187] or by addition of the green fluorescent protein (GFP) sequence to a proteins' coding region, known as GFP tagging [188]. The major drawback to both techniques is that a single experiment must be done for each protein making it time consuming. There is also a chance that the tagging of the protein can interfere with the biology and cause an incorrect result.

The large proportion of known mitochondrial proteins are nuclear encoded and require an N-terminal mitochondrial targeting sequence (MTS) for the protein to be transported into the mitochondrial matrix [189]. Computational efforts to predict mitochondrial protein localisation have been focused on identifying the presence of an MTS for any protein. There have been many different programs which use varying techniques and input data to generate a single score for each protein which represents the probability of the presence of an MTS. The most recent is MitoFates which uses machine learning [33], the other popular programs are iPSORT [190], TargetP [191] and MitoProt [192].

Despite the abundance of different experimental and computation attempts to identify the unknown mitochondrial proteins, there is currently no single study which is widely accepted as the complete mitochondrial proteome although MitoCarta2 is regarded as mostly complete. Multiple different databases have been created which aim to collect various data sets into a single place to help with cross-referencing the data sets on a single protein. MitoMiner [29] contains the most comprehensive set of both experimental and computational mitochondrial protein localisation data sets including fifty different mass spectrometry studies, ten different GFP studies, the four popular MTS prediction programs and the large-scale immunofluorescent study completed by the Human Protein Atlas [187]. MitoMiner references each protein on a gene level and therefore also contains all gene information stored in the gene

ontology (GO) database including GO annotation's on protein localisation [32]. MitoMiner makes cross-referencing the data sets a much simpler task, but the process of identifying whether a protein is localised to the mitochondria remains a difficult task. Many of the data sets contradict each other and there is no method for identifying the most reliable data set without reading each publication, which is extremely time consuming.

4.1.2. Consolidating experimental and computational efforts to predict mitochondrial localisation

When making a judgement on whether a protein is part of the mitochondrial proteome, consolidating the large amount of different mitochondrial protein localisation data sets into a single entity would be vastly better than trying to cross-reference the often-conflicting data sets. Many previous attempts have been made to create a single mitochondrial proteome, with MitoCarta being the most widely accepted [37]. However, the study used only a small subset of the currently available data sets with a relatively simple computational analysis method, a Bayesian classifier. Advances in technology over last decade has seen drastic improvement in machine learning capabilities and availability. Packages for performing machine learning are widely available and continually supported e.g. scikit-learn [97] and Tensorflow [193].

Support vector machines (SVM) is a supervised machine learning technique generally used for binary classification problems. Using a set of samples known to belong to each of the classification groups, called the training set, the method constructs a hyperplane (or set of hyperplanes) in the input variable space which best separates the two classification groups with the largest margin (*Figure 4.1*). Samples of unknown classification can then be assigned to one of the two classes using the generated hyperplane(s). When combined with Platt scaling, SVMs can provide probabilistic class outputs [194]. In the case of mitochondrial protein localisation, this would translate into each protein in the human proteome being assigned a value between zero and one, representing the probability that the protein is localised in the mitochondria. During the training process, the importance of data

sets which have a low amount of classification power are gradually reduced. This means that the final assigned probability will not be majorly affected by any data sets which are lower quality and removes the need to personally read and evaluate each data set. In addition, machine learning methods benefit from having a large amount of input data. All these factors make SVMs the ideal machine learning method for consolidating the vast amount of experimental and computation datasets which try to predict protein mitochondrial localisation into a single, easier to understand entity.

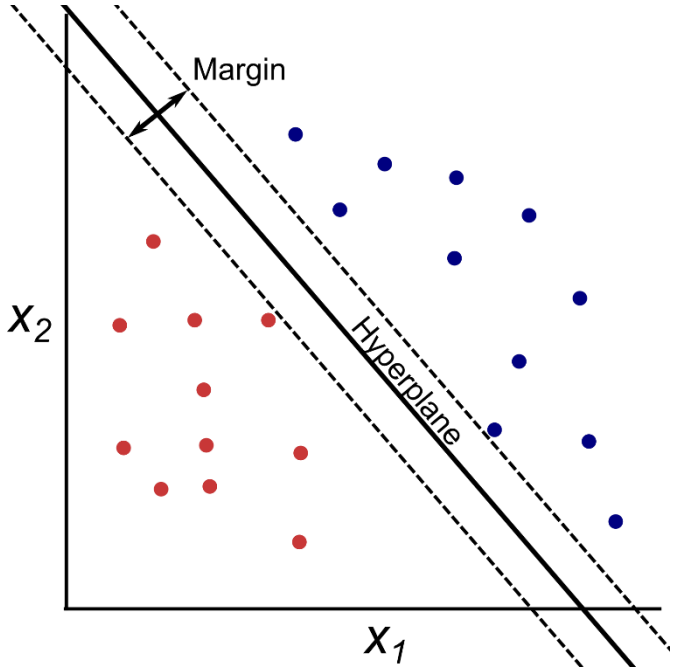


Figure 4.1. An example of the optimal hyperplane which separates the two classes in a two-dimensional variable input space.

4.1.3. Chapter summary

In this chapter, I describe the method for generating and training a support vector machine for predicting protein mitochondrial localisation using a manually curated training set. I outline the iterative tuning process used to generate the input data array and subsequent SVM optimal parameters. The final model generated is then evaluated for overfitting and used to identify novel mitochondrially localised proteins.

4.2. Methods

4.2.1. Mitochondrial protein localisation data source collection

MitoMiner was used as the primary source for collecting mitochondrial protein localisation data sources. The data sources included in the database have already been manually read and checked for any technical flaws. A total of forty-nine different data sources were collected from MitoMiner (*Table 4.1*). Thirty-seven different mass spectrometry mitochondrial protein localisation studies were selected which included a wide variety of different mammalian tissue types [37,183–185,195–227]. An additional large-scale proteomic study using HeLa cells was also selected [31]. The complete set of GFP tagging studies found in MitoMiner were selected, a total of ten studies ranging from mammalian cells to yeast [36,183,188,209,228–233], along with one large-scale immunofluorescence study conducted by the Human Protein Atlas [187]. The final experimental data source selected was a single study which used the enrichment of the CLUH gene, a cytosolic mRNA which specifically binds to a subset of mRNAs encoding mitochondrial proteins, to identify mitochondrial proteins [234]. Four different MTS programs were used to predict MTS on all proteins across four different species; human, mouse, rat and zebra fish [33,190–192]. Homologue cross-referencing across all the different species used in the various data sources was handled by MitoMiner which uses Ensembl Compara [235]. Gene ontology localisation annotations were also taken from MitoMiner for every protein [32]. Two additional studies were collected from outside of MitoMiner which both identified specific species homologs to human proteins, one study for homologs in *Rickettsia* [37] and one study for *Monocercomonoides* [236]. *Rickettsia* homologs were selected as the species is thought to be the most closely related species to the proto-mitochondria. Therefore, its homologs are potentially mitochondrial. In contrast, *Monocercomonoides* homologs are potentially non-mitochondrial as the species has recently been identified as a eukaryote without mitochondria.

Data category	Data	Feature type
Proteomics	37 mass spectrometry studies across various mammalian tissues	Binary, integer
	HeLa subcellular localisation	Binary
GFP tagging	10 studies in mammals and yeast	Binary
Immunofluorescence	Human Protein Atlas anti-body staining	Binary
CLUH enrichment	a cytosolic mRNA which binds to mRNAs encoding mitochondrial proteins	Binary
Computational	Four MTS prediction programs across human, mouse, rat and zebra fish	Unit interval
	Ratio of protein mitochondrial interactors	Unit interval
	COMPARTMENTS annotations	Continuous
Homology	*Rickettsia* homology *Monocercomonoides homology*	Binary

Table 4.1. The data sources collected and used to generate features for the input data array.

4.2.2. Training set definition

The complete set of proteins selected for the positive training set, the known mitochondrially localised proteins, and the negative training set, the known non-mitochondrially localised proteins, can be found in *Supplementary File 4.1*. The two sets were generated and manually curated by Dr. Anthony Smith and Dr. Alan Robinson based on in-depth database and literature reviews. The positive training set contained a total of 1,184 proteins whilst the negative training set contained 1,336 proteins. The set of known mitochondrial proteins (<900 proteins) was expanded upon to generate the positive training set. As part of the iterative process of training an SVM model, proteins which were predicted as mitochondrially localised by a fully trained SVM model were manually literature reviewed. If published evidence for their mitochondrial localisation was found, the proteins were added to the positive training set and the whole SVM training pipeline was re-run.

4.2.3. SVM parameter searching and model selection

All support vector machine modelling was performed using Python [92] and the 'scikitlearn' package [97]. The SVM models were generated using a radial basis function (RBF) kernel. The RBF kernel was the optimal choice as it is one of the most flexible kernels available for SVMs and allows for infinitely complex models. The kernel has also been proven to have universally high generalization ability [237]. All the features in the input data array were normalized to a mean of one and standard deviation of zero before all SVM model generation. To ensure the best choice of kernel parameters *C* and γ, iterative grid searching was performed using a course to fine approach, starting with a large interval for both parameters and iteratively reducing the intervals until the best parameters were identified. A large parameter space for *C* and γ between 0.001 and 10.00 was explored to identify the best possible model using the area under the receiver operating characteristic curve (AUROC) as the metric for measuring model performance. The receiver operating characteristic (ROC) curve is a plot used to illustrate the performance of a classification model by exploring the false positive rate and true positive rate of a model over a variable discrimination threshold. The area under this curve represents

the degree of separability in the model, giving a numeric representation of how well the model can distinguish between the two classes with a value of one being perfect distinguishability. Forty SVM models were generated for each pairwise value for *C* and γ, their mean AUROC was used as the pairwise parameter performance metric.

4.2.4. SVM feature and input data array creation

The input data array which was used in the final SVM model consisted of fifty-seven different columns of information, known as features. The conversion of the data sources into features and the creation of the final data array was an iterative process. The entire parameter grid searching process was performed on many different iterations of the input data array producing many SVM models which were evaluated for performance based on their grid searching AUROC scores and training set accuracies. Additional feature selection was done on each of the input data array iterations using recursive feature elimination on an SVM model generated with a linear kernel, features which performed poorly were ultimately edited or removed from the final data array.

Multiple different types of feature were generated and explored using the MTS data sources. This included using all four programs and each species as a separate feature for a total of sixteen features and using the mean MTS score for each program across all four species for a total of four features. The best performing, and the feature used in the final input data array, was the average score over all four programs and all four species. The single feature was a continuous value between zero and one representing the likelihood of the protein having an MTS based on four different prediction programs across all known homologs in human, mouse, rat and zebra fish. MTS were calculated across four different species as mitochondrial proteins, and their MTS, are expected to be conserved across vertebrates. If a protein MTS is predicted for only one or two species, the result was most likely a false positive.

The mass spectrometry data sources were all converted into binary features representing the presence or absence of the protein in the mitochondrial fraction

analysed in the study. Studies which analysed more than one tissue type were found to generate better models when summed together into a single integer feature rather than separated out by tissue type into multiple binary features. Additional features were created by summing together all binary features of the same tissue type across all the mass spectrometry data sources. A single integer feature was generated in this fashion for placenta, brain, skeletal muscle, liver, kidney and heart. A total of forty-three features were generated from the thirty-seven mass spectrometry data sources and used in the final input data array, a mixture of binary and integer features used in the training of the final SVM model.

A single feature was generated using the data contained in the COMPARTMENTS database [238]. In the database, each piece of evidence supporting either mitochondrial or non-mitochondrial localisation of a protein was assigned a confidence score. These values were collected and the difference between these two values were calculated and used to generate the ‘compartments knowledge difference’ feature where a positive score indicated mitochondrial localisation. An additional feature was generated by Dr. Alan Robinson, labelled ‘Community mitochondrial ratio’, which measured the ratio of proteins that were mitochondrially localised in each protein’s set of direct protein-protein interactors’. The protein-protein interaction information was collected from the STRING database [239].

Eleven binary features were generated from the remaining seven data sources. Two features were generated from the collection of GFP studies, one feature for the collection of mammalian GFP studies and one for the yeast. The GFP studies for each of these two species were combined into a single binary protein presence or absence feature. Four binary features were generated from the HPA immunofluorescence study based on the HPA annotation of protein localisation to the mitochondrial compartment, the non-mitochondrial compartment, not detected and the cytoplasm. Two binary features were generated from the HeLa proteomics study based on the studies’ protein annotation of mitochondrial or non-mitochondrial. A single binary feature was generated from the CLUH enrichment study based on whether a protein’s mRNA had been determined to bind to CLUH during the study. The final two binary features were generated using the homolog

studies in *Rickettisa* and *Monocercomonoides* based on a protein's presence or absence of an identified homolog within the species.

Other types of mitochondrial protein localisation data sources were explored including data sets produced using microarray gene expression [240] and ribosome profiling [241], but these provided no additional classification power and were therefore filtered out during the iterative SVM model creation process. The final fifty-seven feature data array resulted in the best model performance and was used to generate the SVM model for the mitochondrial protein localisation predictions. Hierarchical clustering of the training set over the final input data array and T-distributed Stochastic Neighbor Embedding (t-SNE) [242] plots of the final data array were both generated using R.

4.2.5. SVM model validation

Model validation to check and prevent for overfitting was performed throughout the iterative feature selection and grid searching process. Before every iteration of grid searching, twenty percent of the training set was randomly selected and placed into a held-out validation set with the remaining eighty percent used to train the SVM model. During the training process, randomly selected stratified 10-fold cross-validation was employed to avoid overfitting. The accuracy of the fully-trained SVM model on the held-out validation set was continually compared to the CV accuracy of the model to ensure no overfitting occurred. For each grid searching interval and randomly selected held-out validation set, the remaining training set was split into randomly selected folds (*Figure 4.2*). Eight SVM models were generated for a single training-validation set split, each using a different randomised stratified 10-fold cross-validation. Five different randomly selected held-out validation sets were generated for each grid searching interval, resulting in a total of forty models generated for each grid searching interval. All these steps were to ensure that overfitting was not present, and that the models fully explored the complete training set.

Model validation of the final SVM model produced using the optimal parameters was performed by producing learning curves. Learning curves plot the AUROC error of both the CV and the complete training set against the size of the training set. The curve shows how model performance changes as the training set size increases, giving an insight into the bias-variance trade-off occurring in the model. Learning curves were generated using one-hundred different SVM models generated using the optimal parameters, each with different randomly split stratified 10-fold cross-validation. In addition, overfitting was checked in the final model by calculating held-out validation accuracies and CV accuracies of the SVM model using optimal parameters over a thousand different randomised training-validation set splits.

4.2.6. Input data array feature exploration

The classification power of each of the features used to create the final SVM model was investigated using extreme gradient boosted decision trees. In combination with the positive and negative training sets, extreme gradient boosted decision trees were generated in R using the ‘xgboost’ package [243]. Parameter searching was performed using random grid searching to identify the best parameters based on decision tree accuracy. Using the best parameters, one thousand extreme gradient boosted decision trees were generated, and each features importance value was collected. The feature importance average over all replicates was used as the metric for evaluating overall feature classification power between the positive and negative training sets in the SVM input data array.

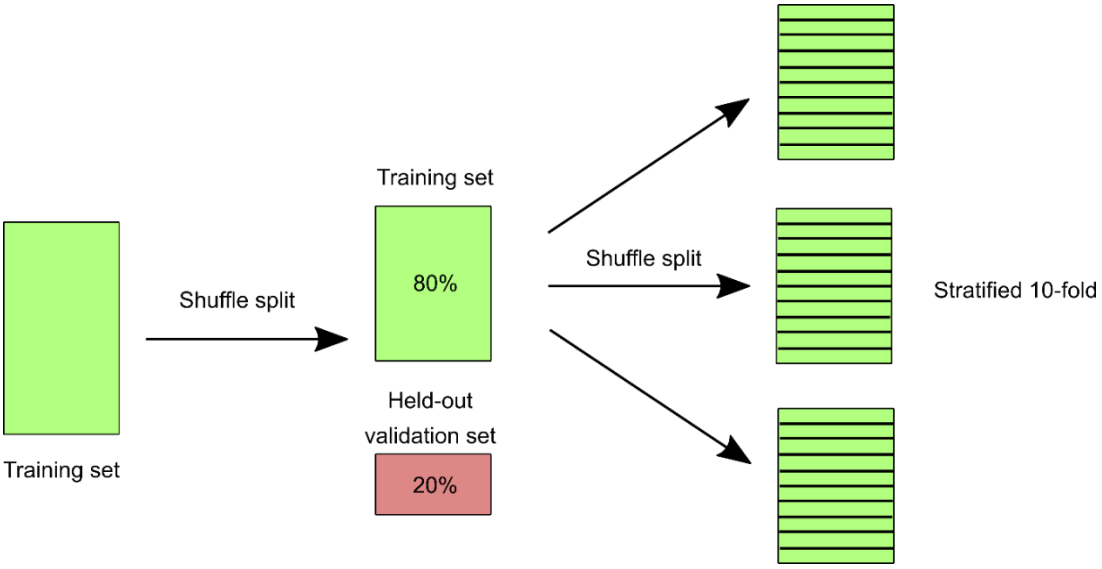


Figure 4.2. An illustration of the random shuffling and splitting of the training set during the grid searching process.

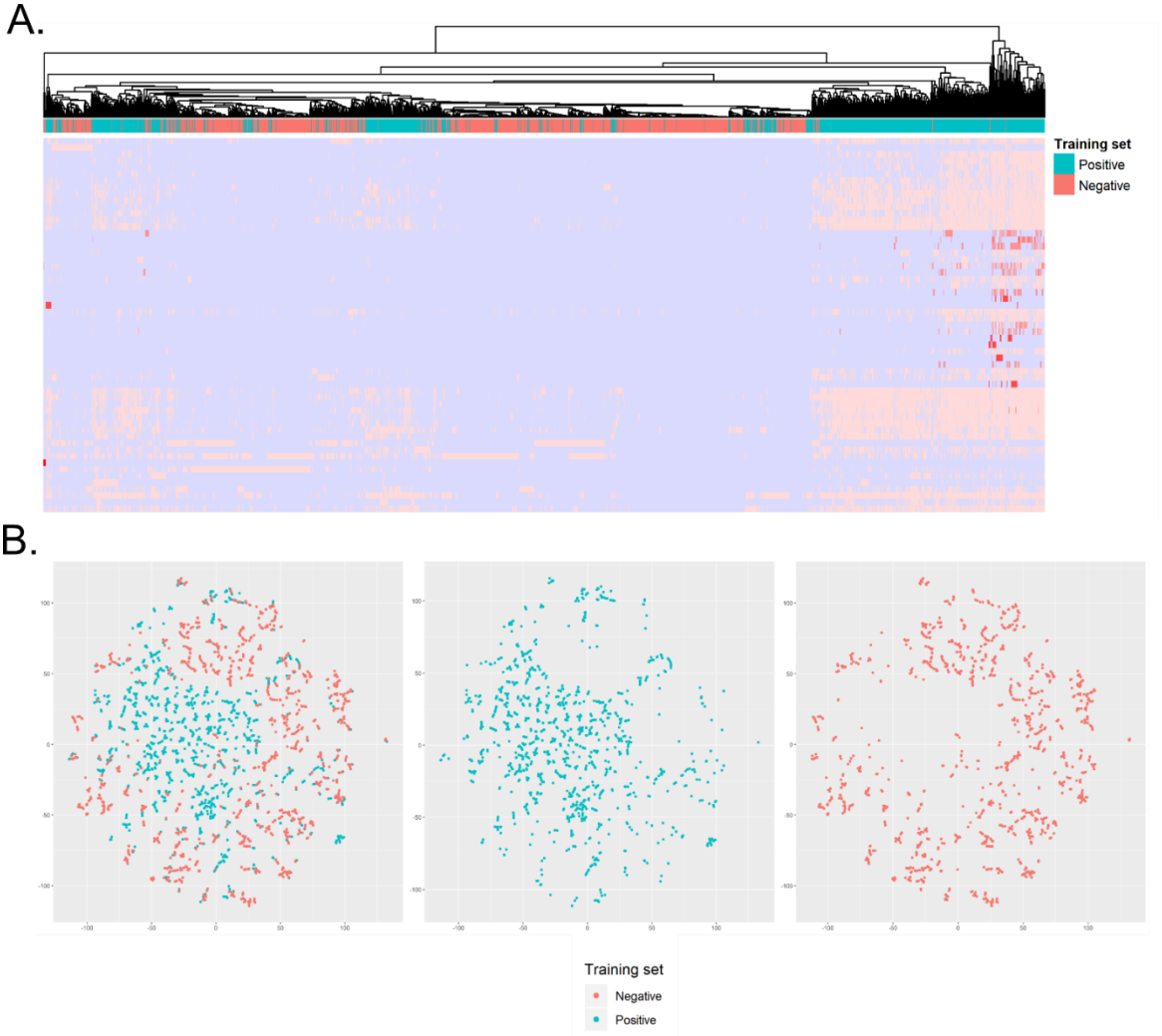


Figure 4.3. A) Hierarchical clustering of the training set samples over the SVM input data array of each protein (column) of the positive or negative training sets, and with each feature (row) scaled to mean zero and standard deviation of one. **B)** t-SNE plot of the training set and input SVM data array with perplexity = 10.

4.3. Results

4.3.1. Training set clustering over the input variable space

Hierarchical clustering of the training set over the input data set showed that the two classes were separable (*Figure 4.3a*). The positive training set clustered into three distinct clusters which were differentiable from the negative training by a large subset of features which were much higher in the positive set. The subset of features included various binary presence or absence features, such as the mass spectrometry studies, which were much higher in the known mitochondrial genes than the known non-mitochondrial genes. The t-SNE plot of the training set over the input array further supports the fact that the classes are clearly separable (*Figure 4.3b*), the two classes showed large regions of clear separation in the plot.

4.3.2. SVM parameter searching using coarse-to-fine grid searching

The grid searching for optimal parameters showed that the overall choice of parameter did not have an overwhelming effect on model performance with only a ten percent difference in the minimum and maximum AUROC scores seen (*Figure 4.4*). The *gamma* parameter for tuning the RBF kernel had the largest effect on model performance with smaller values producing the best performing models. A *gamma* value of 0.0015 was selected as the optimal parameter based on producing high performing models on a consistent basis over the SVM model replicates generated as part of the grid parameter searching process. The *C* parameter for tuning the SVM margin had a much smaller effect on model performance, irrespective of the value of *gamma*. The selection of *C* parameter was set at a value of 2.5 which was selected as a middle ground value. The value was large enough to ensure a small SVM margin such that there was minimal risk of overfitting, whilst being small enough to adequately punish miss-classification and ensure consistently high performing models when paired with the optimal *gamma* value.

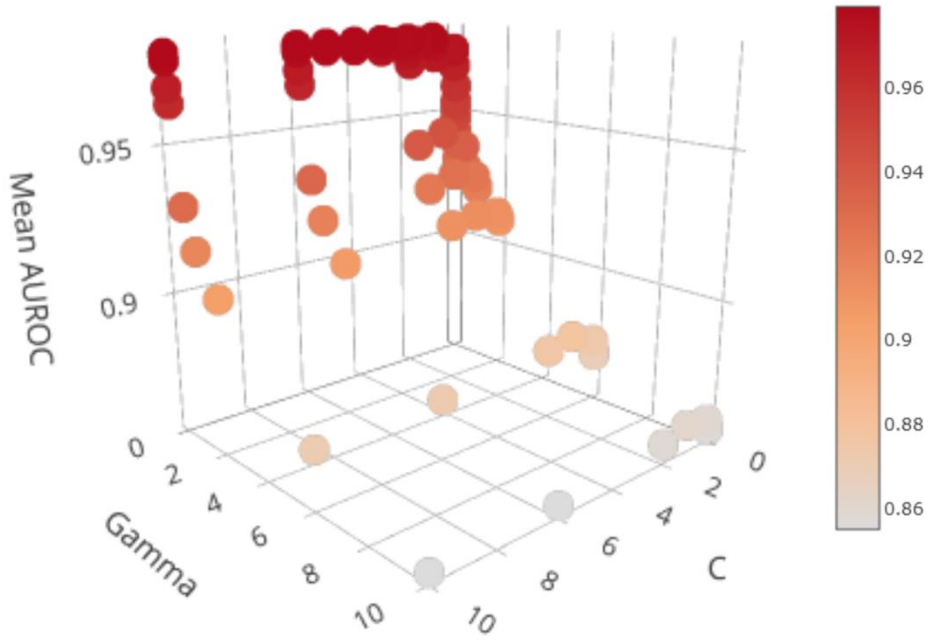


Figure 4.4. A 3-D plot showing the results of the coarse-to-fine grid searching for tuning the SVM model.

4.3.3. Evaluation of the final SVM model

The learning curves produced for the final SVM model showed that cross-validation error consistently improved with increased training set size for almost the entire process of increasing the training set size incrementally (*Figure 4.5a*). The final AUROC error for the cross-validation set eventually reached the low error rate of two percent. As the training set size increased, the training set error and cross-validation error consistently converge to a very similar value, with less than half a percent difference between the final AUROC error values. The cross-validation set accuracy and validation set across the thousand randomly split replicates were both relatively consistent with the sets averaging 92% and 92.2% accuracy respectively (*Figure 4.5b*). The breakdown of the validation set accuracies into the positive and negative set accuracies showed a distinct difference in the class accuracies. The negative set averaged 98.9% accuracy with a markedly smaller variance than the positive set accuracies which averaged 84.7% accuracy.

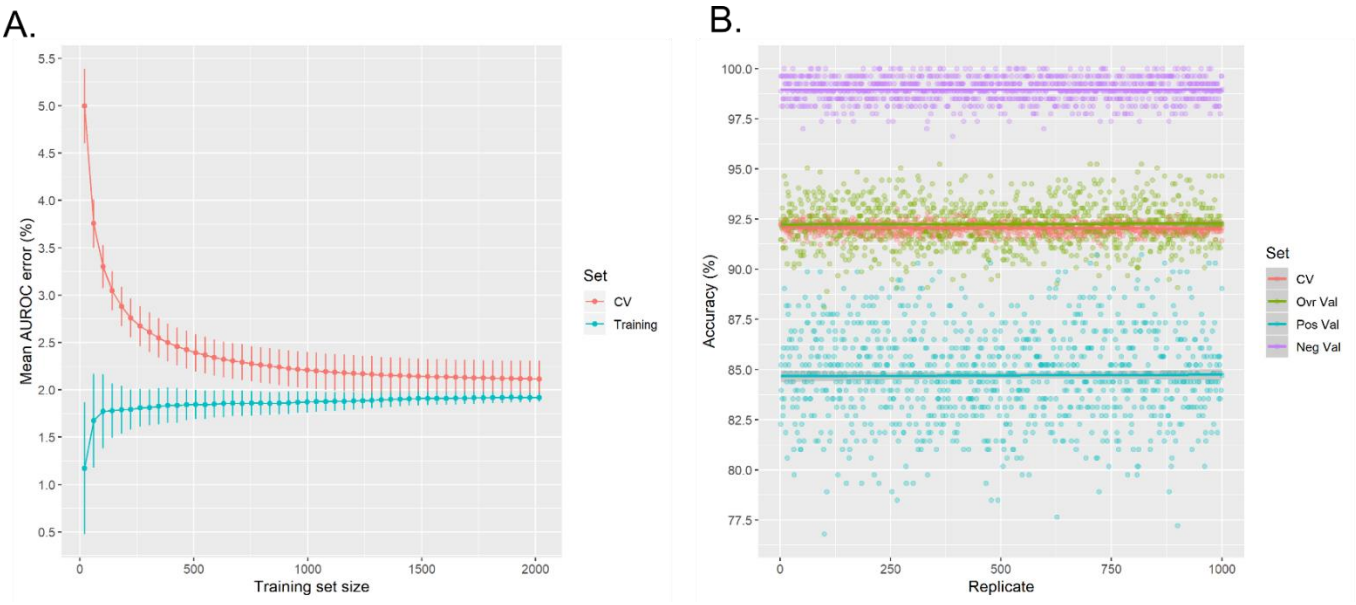


Figure 4.5. A) Learning curve results using the optimal parameterSVM model averaged over 100 replicates. **B)** Accuracies of a thousand replicate SVM models generated using randomly split training sets with optimal parameters.

4.3.4. SVM input feature importance using gradient boosted decision trees

The extreme gradient boosted decision trees identified the COMPARTMENTS annotation feature as the most important feature for classification by a large margin. As such, a second iteration of decision trees were generated with the annotation feature removed to generate more accurate results for the rest of the feature list (*Figure 4.6*). The subsequent trees identified four features as highly important; protein MTS mean score, MitoCarta presence or absence, the protein-protein interaction feature which studied protein interaction clusters and the proteomics HeLa subcellular localisation study. After these top four features, there was a noticeable drop in mean importance score with the second group of important features including the HPA large-scale immunofluorescent study, the study on presence or absence of a *Rickettsia* homolog and a large set of mass spectrometry studies. At the bottom of the list of important features was a large set of mass spectrometry studies, the study on the presence or absence of a *Monocercomonoides sp.* homolog, the CLUH enrichment study and both the mammal and yeast GFP feature. For the training set proteins, each of the important

features were plotted against their final predicted probability for being mitochondrially localised (*Figure 4.7*).

4.3.5. SVM model mitochondrial localisation probability predictions

Upon training the SVM model using the optimal parameters, training set and input data array, the trained SVM model was then used to generate a probability for all known human proteins which indicated their likelihood of being mitochondrially localized. The full breakdown of each proteins predicted probability can be downloaded at http://www.mrc-mbu.cam.ac.uk/impi, the list of novel predicted mitochondrial proteins can be found in *Appendix II*. Of the proteins that belonged to the positive training set, 106 of these were incorrectly classified as they were assigned a probability of less than 0.5 (*Figure 4.8*). In the negative training set, a total of 51 proteins were incorrectly classified with a probability greater than or equal to 0.5 with eleven of these assigned a probability of greater than or equal to 0.8. From the remaining 18,893 known human proteins which were not used in the training set, a total of 1,474 proteins were classified as mitochondrially localized with a probability of greater than or equal to 0.5 with 442 of these having been assigned a value greater than or equal to 0.8.

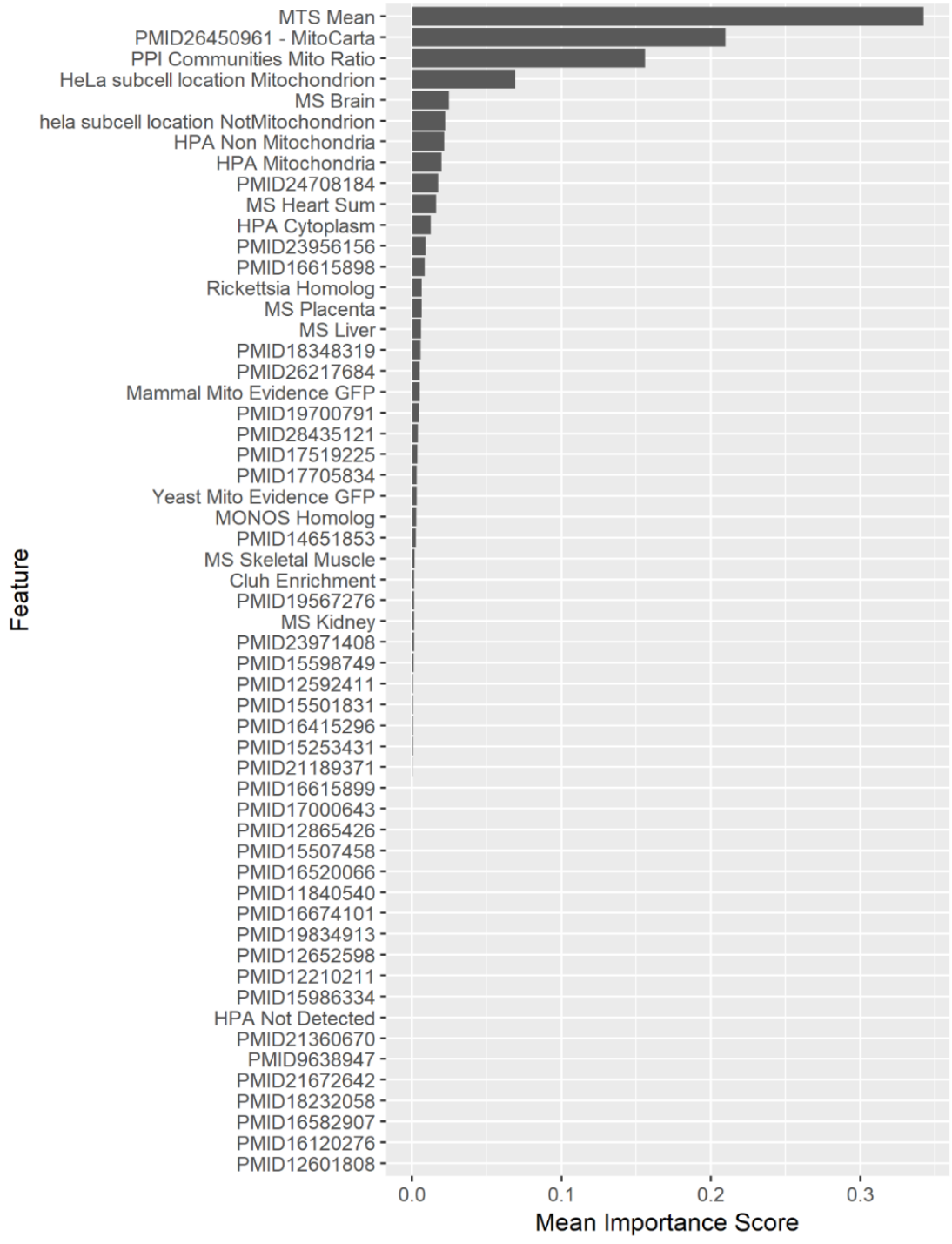


Figure 4.6. The mean importance score for each of the features within the final SVM input data array calculated using extreme gradient boosted decision trees. Mass spectrometry studies are denoted by their PMID, whilst the combined tissue-specific mass spectrometry features are denoted by MS <Organ>.

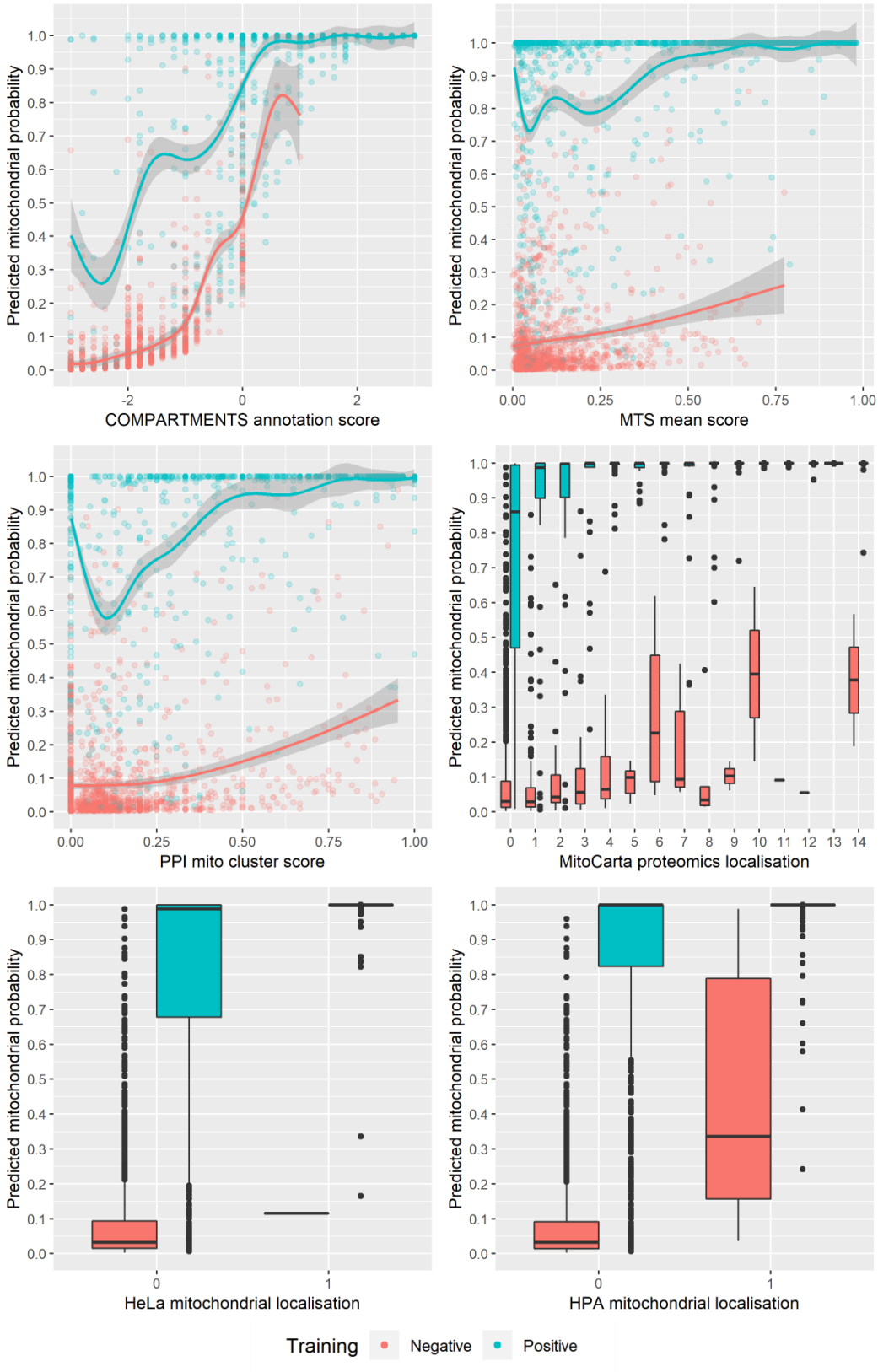


Figure 4.7. Plots of the training set proteins individual feature values against their predicted probability of mitochondrial localisation. Scatter plots have been fitted with a generalized linear model. The MitoCarta graph x-axis represents the number of experiments in which a protein was identified as mitochondrial during the study.

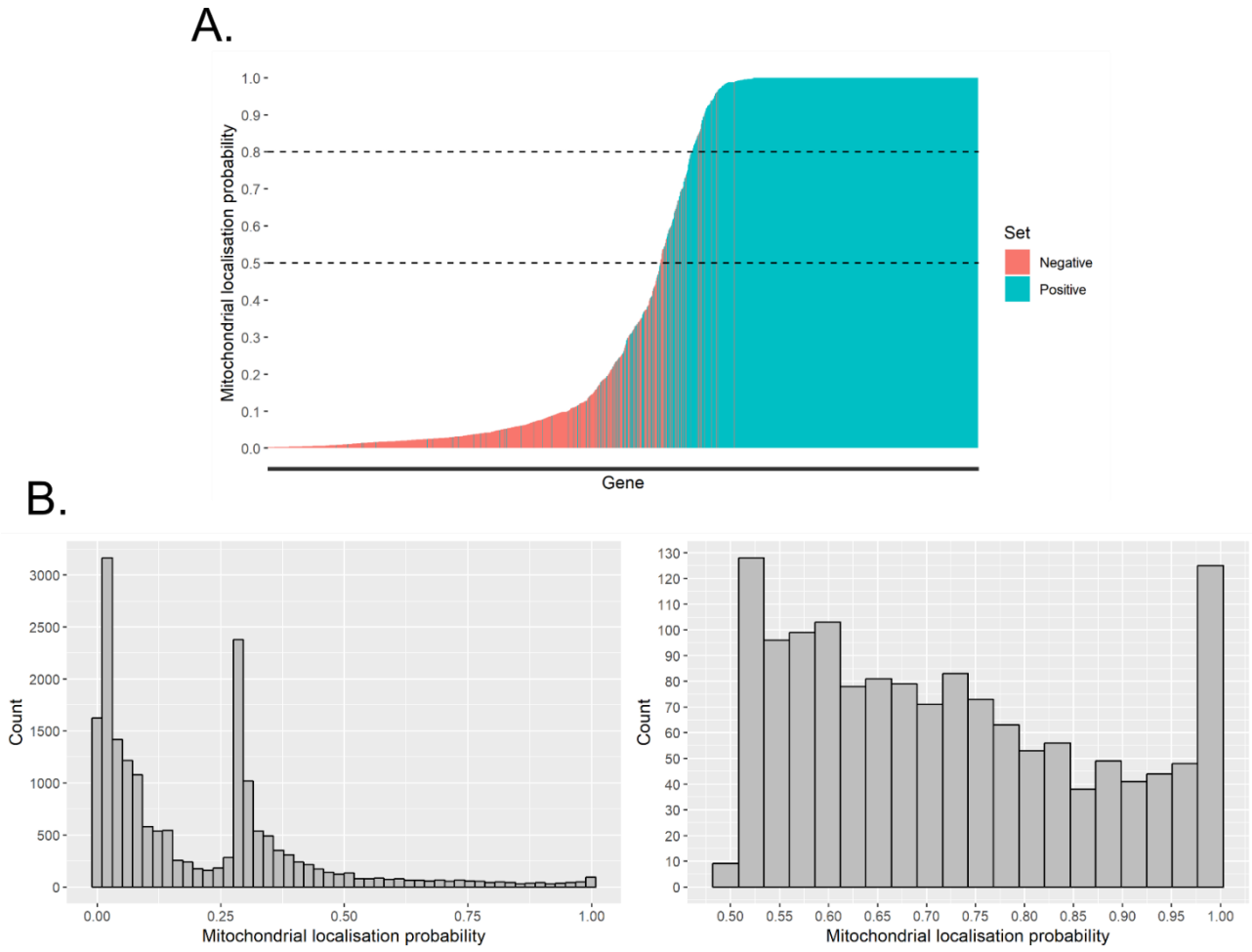


Figure 4.8. A) The predicted mitochondrial localisation probabilities of the proteins belonging to the positive and negative training sets. **B)** Histogram's of the predicted mitochondrial localisation probabilities of the other 18,893 known human proteins which were not included in either training set.

4.4. Discussion

4.4.1. Validation of the final SVM model

The performance of any machine learning model is entirely limited by the quality of the input data array and the selection of the training set, the training set must be representative of the two classes and be separable in the input variable space. The t-SNE plot and hierarchical clustering of the training set confirmed that this was true for the generated input array and selected training set (*Figure 4.3*). Both methods vastly simplified the fifty-six-dimensional input array and managed to clearly separate the two classes. This highlighted the viability of the input array and training sets for use in more complex algorithms such as SVMs using the RBF kernel. The slight overlap of the training sets in the t-SNE plot, and presence of sparsely spread out positive training set samples within the negative training set clusters in the hierarchical clustering, indicated the presence of potential positive set outliers which could negatively affect model performance. However, as the training set was manually curated, these were experimentally confirmed mitochondrially localised proteins, not outliers, and were therefore kept in the positive training set to ensure complete representation of mitochondrially localised proteins within the training set. To ensure that the entirety of the positive training set was considered within the training process, including those that clustered with the negative set, the iterative random splitting described in *Figure 4.2* was performed for many replicates during the grid searching process.

The complete representation of mitochondrially localised proteins within the training set and the performed grid searching method was employed to prevent overfitting. The final model showed no sign of having overfitting, also known as high variance, based on the results of the learning curves and repeated held-out test set accuracies (*Figure 4.5*). In the presence of overfitting the training set error would remain extremely low whilst the cross-validation error would not improve resulting in a large difference in error rate. Both the training and cross-validation curves continually converged and eventually ended with only a small difference in error rates, an indicator that overfitting was unlikely. A large difference in held-out validation accuracy and cross-validation accuracy would have been seen if the model had been overfit to the training set. Many replicate SVM models were

generated using randomly selected held-out validation sets to ensure that all training set samples were included in the held-out validation sets to improve the reliability of the accuracies. Only a 0.2% difference was seen in the average accuracy of the cross-validation set compared to the held-out validation sets with both having a relatively small variance, further indicating that the model did not suffer from having high variance.

The flip side of a model overfitting is the model underfitting, also known as high bias, in which the algorithm fails to identify important aspects of the input variable space for the classes. An obvious indicator for underfitting would be a high overall error rate for both the training and cross-validation sets which was not present in this study. Additionally, the gradient of the learning curve for both the cross-validation and training sets would not gradually decrease until almost zero, which is precisely the behaviour present in the model learning curves. Both clearly show that the model did not suffer from having high bias. Combined, these results indicate that the final model had a good balance in the bias/variance trade-off and did not suffer from either under or overfitting.

The extreme gradient boosted decision trees identified six features which would be expected to have a large amount of classification power in the final SVM model; three continuous, two binary and one sum of binary variables (*Figure 4.6*). The plots showed that none of these features singularly dominate the SVM probability predictions (*Figure 4.7*). In each of the features, both positive and negative training set samples are almost entirely classified correctly using a 0.5 probability threshold, irrespective of a single feature value. For example, using the generalized linear model as an indicator, positive training set samples with both a low and high MTS mean score are predicted to have a probability over 0.7. The predictions are therefore not being made by the algorithm overfitting to a single feature but by a combination of multiple features. This result supports the fact that no overfitting occurred in the model and gives confidence to the choice of SVM model kernel, tuning process and selected parameters.

4.4.2. Mitochondrial proteome predictions of the final SVM model

The average held-out validation set accuracy of 92.2% is the generalisation accuracy of the final model, the expected accuracy on the prediction of mitochondrial localisation for a completely unknown protein (*Figure 4.5b*). The large difference in the average validation accuracy between the positive and negative training set samples, and the difference in variance of the accuracies over the replicates for each set, identifies a potential way for increasing the model accuracy. The lower validation accuracy of the positive training set samples, and larger variance over the replicates, would most likely have been caused by the positive training set samples which were seen overlapping with the negative training set samples in the t-SNE plot, and those which clustered with the negative set in hierarchical clustering. By removing these samples, a higher positive set validation accuracy would most likely occur, but this would come at the cost of removing confirmed mitochondrially localised proteins and reduce the positive set representation of mitochondrially localised proteins. As such, and due to the already high generalisation accuracy of the final model, the positive training set was not altered. The difference in each training set validation accuracy identified a much higher model specificity than sensitivity i.e. proteins that are predicted to be non-mitochondrially localised (predicted probability less than or equal to 0.5) can be taken with very high confidence but those which are predicted to be mitochondrially localised (predicted probability greater than 0.5) should not be taken as absolutely certified.

In the case of identifying mitochondrially localised proteins, having a higher specificity is much more desirable than a high sensitivity. Outside of the two used training sets, the human proteome consists of 18,893 proteins which must be investigated when trying to identify novel mitochondrially localised proteins. The final model predicted that only 1,474 of these proteins had a probability of being mitochondrially localised, leaving 17,419 proteins predicted non-mitochondrial which, due to the hyper specificity of the model, can be taken with high confidence. In cases where novel mitochondrially localised proteins need to be identified, such as for the investigation of serious mitochondrial diseases, the list of potential proteins can therefore be confidently reduced to 1,474 proteins. For defining a

mitochondrial proteome with confidence, based on the probability predictions of the final SVM model, the hyper specificity means that utilizing a higher probability threshold for class assignment would be beneficial.

Using a probability threshold value of 0.8 instead of the default 0.5, the number of false positives from the training set was reduced from 51 to just eleven out of the 1,336 proteins belonging to the negative training set (*Figure 4.8*). This new threshold value was selected solely for defining a referenceable mitochondrial proteome and was selected based on manually evaluating predicted proteins at various probability values on their evidence of protein localisation, their protein function and in-depth literature reviews. Using a probability threshold of 0.8 identified 442 predicted mitochondrially localised proteins from the 18,893 unknown human proteins. Combined with the known mitochondrial proteins used in the positive training set, generated a predicted mitochondrial proteome of 1,626 proteins, a number close to the expected 1,500 proteins of the mitochondrial proteome. The remaining 1,032 proteins which were assigned a probability between 0.5 and 0.8 should still be considered when trying to investigate novel mitochondrial proteins as the proteins still show some evidence for being mitochondrially localised, just at a lower confidence level.

4.5. Conclusion

The final SVM model generated using the created input data array along with the training set performed very well and showed no sign of over or underfitting the data, with a generalisation accuracy of 92.2%. The model predicted 1,474 proteins to have adequate evidence to suggest mitochondrial localisation, with 442 having very high evidence. Therefore, the SVM model identified 442 novel mitochondrially localised proteins with high confidence, generating a predicted mitochondrial proteome of 1,626 proteins. The predicted proteome is currently being used in an NGS analysis pipeline within the MRC Mitochondrial Biology Unit for diagnosing mitochondrial disease patient samples.

Chapter 5

Predicting mitochondrial disease

5.1. Introduction

5.1.1. Genetic causes of mitochondrial disease

The complex function of the mitochondria and its highly interconnected network of pathways leads to mitochondrial diseases having many different clinical phenotypes. Patients with mitochondrial disease caused by in-born errors often present with unique clinical phenotypes which are often difficult to associate with the precise cause of the disease. For example, patients with a mutation in the mitochondrial citrin transporter, SLC25A13, can present with neonatal intrahepatic cholestasis or adult onset hepatic disease, neither of which have an obvious correlation to citrin transportation across the mitochondrial matrix [244]. As mitochondria have their own DNA, called mitochondrial DNA (mtDNA), mitochondrial disease can be caused by mutations in either the mtDNA or nuclear DNA (nDNA).

In humans, mtDNA are small circular DNA present in the mitochondrial matrix which are fully defined and encode for thirty-seven genes [27]. In humans the mitochondrial genome is entirely maternally inherited [245]. The mutation rate of the mitochondrial genome is much higher than the nuclear genome, believed to be related to the oxidative stress caused by OXPHOS within the mitochondrial matrix [246]. However, in a study of patients with pathogenic mtDNA point mutations, de novo mutations were common but showed low reoccurrence in future generations [247]. The presentation and clinical phenotype of mitochondrial diseases caused by mtDNA mutation are complicated by mitochondrial heteroplasmy, the proportion of wild-type mtDNA in the mitochondrial population compared to the mutant mtDNA.

Heteroplasmy of mutant mtDNA must exceed a threshold before mitochondrial diseases manifest themselves in patients [46], with a higher proportion of mutant mtDNA having been associated with more severe symptoms [248]. Cases of mitochondrial disease caused by mutations in mtDNA have been estimated at 9.6 cases per one-hundred thousand people [249]. Examples of mtDNA mutations which result in mitochondrial disease include mutations to complex subunits such as MTND1 and MTCO1 for complex I and IV respectively [250].

The human nuclear genome shows Mendelian inheritance and encodes for around 21,000 proteins. There are 1,184 known proteins which belong to the mitochondrial proteome with an extra 442 predicted to localise to the mitochondria by the support vector machine trained in the previous chapter. The much larger library of possible nuclear encoded proteins causing mitochondrial disease makes diagnosis of nuclear causing mitochondrial diseases a more difficult process than diagnosing those caused by a mutation in the mtDNA. Cases of mitochondrial disease caused by mutations in nDNA have been estimated at 2.9 cases per one-hundred thousand people [249]. Examples of nDNA mutations which result in mitochondrial disease include mutations to the nuclear encoded respiratory complex subunits, such as NDUFS1 for complex I, and complex assembly factors such as SURF1 for complex IV [250,251].

5.1.2. Diagnosis of mitochondrial disease by next generation sequencing

Genetic screening for mitochondrial disease is the primary method for identifying and diagnosing mitochondrial disease. As mtDNA is relatively small, any patient with a suspected mitochondrial disease—especially if a family history suggests maternal inheritance—will have their mitochondrial genome sequenced first, generally using next generation sequencing (NGS) [252]. If the mtDNA lacks a plausible pathogenic variant, the patient will have their nuclear genome sequenced by either targeted NGS [253], the sequencing of a selected panel of genes, whole exome sequencing by NGS [254] or whole genome sequencing by NGS [255]. Recent studies have highlighted whole exome sequencing as the superior method for diagnosing mitochondrial diseases [255–257]. The sequence data collected from suspected

mitochondrial disease patients is then put through a pipeline such as the Genome Analysis Toolkit (GATK) [258]. The pipeline compares a patient's genome with a reference genome and identifies all the mutations present in a patient, known as variant calling. Studies on the genetics of the human population shows that all humans carry an abundance of rare variants [259,260]. Therefore, each patient usually has hundreds of variants in the genes belonging to the mitochondrial proteome which must be filtered through in order to identify the disease-causing variant. When possible, family members are usually put through a similar pipeline to enable additional filtering of a patient's variant list.

Initial screening of the potential variants is done by comparison to the already known mitochondrial disease genes. The Online Mendelian Inheritance in Man (OMIM) database stores information regarding all human disease genes and their publications [261]. For mitochondrial diseases specifically, a manually curated and constantly maintained list of mitochondrial disease genes has been generated by a group at Washington University [262]. Variants which occur in one of the 344 known mitochondrial disease genes are then filtered by their rarity in the population. The Exome Aggregation Consortium (ExAC) database contains whole exome sequences from 60,000 healthy individuals and their known variants which allows for each variants rarity within the population to be calculated [263]. Variants in known mitochondrial disease genes which are rare within the population (<0.001%) are most likely to be the cause for disease. In cases where these variants do not exist, patients are left with a huge list of variants which may contain the disease-causing variant. This list must be manually investigated which is extremely time consuming and often almost impossible to filter.

Large-scale studies on the diagnosis of mitochondrial diseases using sequencing identified only a 50% success rate [264,265]. Novel dominant disease-causing variants (those which are caused by a single mutation in one allele) are particularly difficult to distinguish from non-pathogenic variants in genes not associated with a dominant disease, unless the sequences of both parents are known. Furthermore, some diseases may be caused or exacerbated by the compound effect of multiple mutated genes leading to different penetrance [266]. Thus, the list of single allele variants is usually huge for each patient. The largest hurdle in diagnoses of patients

with mitochondrial disease is this inability to identify candidate disease-causing genes from a patient's list of variants which can be experimentally studied and verified.

5.1.3. Computational methods for identifying disease genes

Experimental procedures for identifying disease-causing genes are extremely costly and time consuming. Reducing the list of potential disease-causing variants in a patient to a smaller set of potential candidate genes using computational methods would enable a more targeted experimental approach and increase the diagnosis rate of mitochondrial diseases. The wealth of high quality, manually curated databases being created such as Gene Ontology (GO) and UniProt has seen the rise in computational methods being explored that can be used in conjuncture with sequencing to filter a patient's list of variants. For example, exploring the functional annotation of a protein and linking its potential loss of function to the patient's symptoms [267]. An alternative approach focuses on predicting potential disease genes which can be used to improve the screening of a patient's variants.

Many computation methods for predicting disease causing genes focus on studying the protein-protein interaction network of the human proteome [268,269]. Using network analysis, the local information in a protein-protein interaction network can be explored and used to quantify the positional importance of a protein within a local cluster [270]. Local clusters of the human protein-protein interaction network have been shown to fulfil similar cellular functions [271]. Within these clusters, disease genes have been shown to segregate at the edge of clusters and avoid highly connected areas [272]. In addition, disease genes have been shown to have a high propensity to interact with each other, forming disease-related clusters of proteins [272]. A method for predicting disease genes using this theorem involves comparing the interactors of a protein to those of a known disease gene for similarity [273,274]. The global features of the human protein-protein interaction network such as its topology have also been explored to predict disease genes. A study which used all known disease genes in OMIM showed that human disease genes had a larger number of interactors, had a shorter distance to other disease genes and shared

more common interactors to other disease genes than genes which had no disease association [275]. These theorems identify a potential method for predicting mitochondrial disease genes using the human protein-protein interaction network which has not yet been explored.

The increased availability in high throughput gene expression data has seen the increase in studies trying to utilize these data sets in identifying disease genes. The recent discovery of a tissue-specific gene thought to be a major contributor to type 2 diabetes portrays the importance of studying tissue-specific expression when considering disease genes [276]. In general, disease genes have been identified as overexpressed in tissues with the highest pathology but the relationship between tissue specific expression of a gene and the tissue pathology was not the same for all diseases [277]. In addition, the upregulation of disease genes in their affected tissues has higher association with autosomal dominant diseases compared to recessive [278]. The tissue-specificity of mitochondria due to variable tissue energy demands highlights the importance of exploring tissue-specific expression when trying to predict mitochondrial disease genes and may be a potential way to improve dominant disease gene identification.

For mitochondrial disease genes specifically, a recent study has identified a link between mitochondrial gene evolution and their association with disease [279]. The study identified that the known mitochondrial disease genes were more likely to have orthologue's in a wider set of evolutionary taxa than those which are currently not associated with disease. In addition, the origin of the known mitochondrial disease genes was more likely to be early in evolutionary history, before the evolution of eukaryotes, than those which are currently not associated with a disease. These findings suggest that using a gene's evolutionary history could be a method for identifying novel mitochondrial disease genes.

5.1.4. Neural networks for predicting mitochondrial disease genes

With the rise of machine learning over the last few years, there has been many attempts to use machine learning with different data sources to try and predict

disease genes in hopes of developing a method for precision medicine [280]. A random forest classifier was trained using protein function similarities to try and predict autism spectrum disorder genes. The functional similarities of every protein were assessed using their GO annotations and protein-protein interactors. The model had moderate success with a reported AUROC of 0.8 [281]. There has also been an attempt to predict Parkinson's disease genes which was done by training a support vector machine using only protein-protein interaction network information. The SVM was trained on protein positional importance in the human protein-protein interaction network which was quantified using random walks across the network. The model also had moderate success with a reported AUROC of 0.73 [282]. Tissue expression data has also been used as a feature of machine learning, it was exclusively used to train a random forest classifier to diagnose cancer [283]. The classifier managed to distinguish each of the cancerous tissues from their healthy counterparts.

Only a single attempt has been made to use machine learning to specifically predict mitochondrial disease genes [284]. The study used a broad range of data sets for training which included protein ortholog and gene expression data. However, the purpose of the machine learning algorithm was to predict protein function as an inference for mitochondrial localisation. The identification of disease genes was then done after experimental verification of a patient's disease locus. Candidate prioritization of variants was done by assessing the predicted mitochondrial localisation score of the genes known to be around the disease locus. The method did not directly address the issue of predicting mitochondrial disease genes and is built around experimental data which requires a large amount of time and cost investment.

Most machine learning models trained for disease gene prediction used relatively simple machine learning algorithms such as support vector machines and random forests. Neural networks are a supervised machine learning algorithm which can be trained for classification problems [285]. The networks are made up of multiple connected layers of nodes, called neurons, which train in an iterative manner like other supervised machine learning algorithms. However, neural network structure can be manipulated by changing the number of layers and nodes in the network

which allows for more complex relationships between the inputs to be identified and used for classification, known as deep learning. The final output of a neural network for each input is a vector of class assignment scores which can be used to assign a class to each input. A recent study which tried to predict disease genes used a deep belief net, a class of neural network [286]. The neural network was trained using GO annotation data and a latent representation of the human protein-protein interaction network. The model had a reported AUROC of 0.97, much higher than the machine learning models generated in a similar manner which used simpler machine learning methods. This highlights the potential for using a neural network for predicting mitochondria disease genes using the theories on mitochondrial disease genes mentioned in the previous section. Neural networks could be trained using the same input information for predicting disease association and disease type, such as dominant or recessive. These neural networks would provide a score for disease association, and for potential disease type, which could be used to prioritize a patient's variants for experimental investigation and greatly improving the diagnosis rate of mitochondrial diseases.

5.1.5. Chapter summary

In this chapter, I discuss the process of collecting and analysing features which can be used to differentiate known mitochondrial disease genes from suspected non-disease mitochondrial genes. I describe the complete training process of two neural networks using the complete set of features as the input data array, one network for binary classification of disease or non-disease, and one for multiclass classification of non-disease, recessive disease, dominant disease or other disease. I then examine the trained networks for over-fitting and evaluate their performance on the training set. The final trained neural networks are then used to predict novel mitochondrial disease genes, the resulting predictions are compared between the two networks and investigated.

5.2. Methods

5.2.1. The mitochondrial disease gene training set

The previous chapter identified an extended mitochondrial proteome of 1,626 proteins, each of which has the potential for being a mitochondrial disease gene. The predicted proteome was therefore the test set for both the trained neural networks, the set of proteins which were eventually put through the trained neural networks to generate novel predicted mitochondrial disease genes. The set of 344 known mitochondrial disease genes, as defined by the OMIM database and Wash U mitochondrial disease gene list, were used as the positive training set for both neural networks. Their disease classification, such as dominant or recessive, were collected from MitoMiner and separated into three categories; dominant, recessive and other. The other classification group contained all other disease classes such as X-linked dominant and X-linked recessive. The positive training set contained 20 dominant disease genes, 273 recessive disease genes and 51 other disease genes. The negative training set, the set of known non-disease genes, was generated from studies on loss of function mutations. Four different population studies identified 214 genes in the predicted proteome which had rare (<1% minor allele frequency) loss of function (LoF) homozygote mutations present in healthy individuals [263,287-288], i.e. both alleles are expected to be non-functional. As these LoF mutations were present in healthy individuals, their likelihood of causing disease is extremely low, making them the best choice for ‘known’ non-disease genes. The negative test set was therefore made up of 214 genes which were used to train both neural networks. The set of genes used in the positive training set, along with their disease association, and the negative training set can be found in *Supplementary File 5.1*.

5.2.2. Mitochondrial disease gene neural network input data array creation

The complete data input array used to train both neural networks can be found in *Supplementary File 5.2*. Seven input array features were generated using a human protein-protein interaction network (*Table 5.1*). The protein-protein interaction network was created by combining protein interaction data from multiple databases, each of which were themselves a combination of many different databases;

inBioMap [290], ConsensusPathDB [291], mentha [292], Integrated Interactions Database (IID) [293] and STRING [239]. The inBioMap and STRING databases include confidence scores with each of their protein-protein interactions indicating the level of confidence in the evidence supporting the interaction. For inBioMap interactions, only those with a mean confidence interval greater than 0.1 were included whilst only the interactions with a score greater than 0.9 from the STRING database were included. The combined protein-protein interaction network consisted of 20,408 nodes (proteins) with 1,148,406 edges (interactions). All network exploration and the calculation of the network metrics which were used as features were generated in R using the 'igraph' package [294], unless stated otherwise.

Five different features were generated using the complete protein-protein interaction network. For each of the predicted mitochondrial proteome proteins their number of interactors (degree), clustering co-efficient (transitivity), number of shortest paths through the protein (betweenness centrality), and number of interactions required to completely disconnect the protein from each of the known mitochondrial disease genes (edge connectivity) were calculated. The edge connectivity of a protein to all known mitochondrial disease genes was averaged and used as a single feature, the other four metrics were each used to generate their own single features. The final feature generated using the complete protein-protein interaction network involved the use of self-avoiding random walks which were completed using custom scripts in MATLAB. For each protein in the complete interaction network, five hundred self-avoiding random walks of length 1 to 10 were completed. Using the known start and end point of the complete set of random walks over the network, *outward accessibility* as described by *Travençolo et. al* [295] was calculated. Outward accessibility is a network metric which quantifies the positional importance of each protein in the network by evaluating the accessibility of each node in the network with respect to all other nodes. A single feature in the input data array was generated using each protein's calculated outward accessibility.

Two additional protein interaction features were generated using clusters of the protein-protein interaction network. Overlapping clustering of the network were calculated using overlapping cluster generator (OGC) [296]. For each cluster in which a predicted mitochondrial protein was present, the ratio of known

mitochondrial disease proteins vs unclassified mitochondrial proteins was calculated and assigned to every predicted mitochondrial protein in that cluster. The average of this ratio for each protein was then used to generate a single feature. In addition, for each cluster in which a known mitochondrial disease gene was present, the structural similarity, a measure of common interactors between two nodes in a network, was calculated between the disease gene and all other predicted mitochondrial proteins within the cluster. The average of this value for each protein was then used to generate a single feature.

Tissue specific expression of each gene was taken from the Human Protein Atlas RNA-Seq study, collected from MitoMiner. The study measured the expression of each protein in thirty-seven different tissues, each tissue type was used to generate a single feature. The sequence length and protein mass for each protein were collected from the UniProt database and used to generate two features. In addition, UniProt annotations for protein function and pathway activity were collected for each protein. Using the training positive and negative test sets, each of the annotation labels were assessed for classification power using χ^2 testing. Only two annotations showed a significant difference between the positive and negative training sets ($p = 3.10x10^{-6}$, $p = 0.001105$), the ‘acetylation’ and ‘transport’ functional annotations, which were used to generate two features. The transport annotation being significant was unsurprising due to the impermeability of the inner mitochondrial membrane forcing a high importance onto the transportation proteins to ensure regular mitochondrial function. A potential reason for the significance of the acetylation annotation was that acetylation has been identified as a regulatory mechanism for mitochondrial proteins [297,298], dysfunction of which would cause a large disruption to regular mitochondrial function.

Two features were generated using the population genomic study available on the ExAC database. The database has its own calculated metric for each gene called the probability of being loss-of-function intolerant (pLi) [299], along with z-scores on the number of observed vs. expected missense mutations identified in each of the genes. Each metric was used to generate a single feature and used in the input data array.

The final set of twelve features used in the input array was generated using the results of the mitochondrial protein evolutionary study [279]. In the study, each protein's homologs were identified across various phylogenetic taxa. Eight different features were generated based on the number of homologs found in the eight studied taxa for each protein, along with a single feature based on the total number of homologs found across all eight taxa for each protein. In addition, a single feature was generated using the number of unique taxa in which each protein had identified homologs. Finally, a categorical feature was generated for each protein based on the protein's oldest taxa in which a homolog was identified. Each of the eight taxa were assigned a value between one and eight as a class assignment which was used to generate a single categorical feature. Each of the eight taxa was assigned an estimated age since their first evolutionary appearance in millions of years. Using the categorical class assignment for each protein, a single feature was generated to represent a protein's estimated age since its first appearance in evolution.

Exploration of the input data array and training set selection prior to developing the neural networks was performed using R. Each of the features created using the protein-protein interaction network, and those generated from the evolutionary study, were tested for significance over the training sets using Wilcox's rank sum testing [300]. The tissue expression features were explored using hierarchical clustering across the training set. The entire input data array was used to train an extreme gradient boosted decision tree using the 'xgboost' package [243]. Parameter searching for the best decision tree was performed by generating three hundred randomly initiated gradient boosted decision trees. The best performing parameters were then used to generate five thousand replicate gradient boosted decision trees. Importance scores for the features were extracted from each replicate and averaged. The entire input array was also used to generate a t-SNE plot.

Data category	Data	Feature type
Protein information	Protein mass	Integer
	Sequence length	Integer
Protein interaction network	Degree	Integer
	Transitivity	Unit interval
	Betweenness centrality	Continuous
	Edge connectivity	Continuous
	Outward accessibility	Continuous
	Cluster mitochondrial disease ratio	Unit interval
	Cluster structural similarity	Unit interval
ExAC population studies	Z-score for missense	Continuous
	probability of being loss-of-function intolerant (pLi)	Unit interval
HPA tissue expression	RNA-Seq data for 37 tissues	Continuous
UniProt annotation	Acetylation	Binary
	Transport	Binary
Evolutionary study	Identified homologs in eight taxa	Integer, Categorical

Table 5.1. The data sources used to generate features for the input data array.

5.2.3. Neural network creation, tuning and validation using TensorFlow

Two neural networks were trained using the same input array and training set with TensorFlow in python [193]. One network was generated for binary classification, predicting disease or non-disease, whilst the other utilized the positive training set split into their disease types. The split positive training set was used to train a neural network for multiclass classification, predicting non-disease, dominant disease, recessive disease or other disease. To reduce the chance of overfitting and allow for accurate evaluation of the neural networks during the tuning process, the training data for both networks was split into a training set, validation set and test set. The validation and test set each consisted of ten percent of the training set and were separated from the training set in a stratified manner (the ratio of positive and negative training set samples in each split was kept the same across all of the splits). The seed used to split the data into the three sets was kept the same throughout the generation of both neural networks to enable comparisons between different networks whilst tuning the hyperparameters.

Training of both neural networks was done using batch gradient descent [301] with the Adam optimization algorithm to improve the learning speed [302]. All hidden layers used in the network architectures were dense layers, the inputs were batch normalized [303] using batch renormalization due to the potentially small batch sizes [304]. Drop-out normalization was employed just before the leaky rectified linear unit (ReLU) activation function [305,306] of each layer to improve training performance and reduce the chance of overfitting [307]. During training the loss was calculated using sigmoid with cross entropy for the binary neural network, and softmax with cross entropy for the multiclass network. Bias variables were initiated as zero, whilst the weights for each neuron were randomly initialised using He initialization [308]. Optimal parameters for the network were identified using hyperparameter searching. Six different parameters were investigated using randomly selected values between specific ranges in a coarse-to-fine grid searching methodology; learning rate of the network between 0 and 1.0, batch size of the input size during training between 5 and 200, the number of nodes in each hidden layer with one variable for each hidden layer between 2 and 500, and the drop-out probability for each layer with one variable for each layer between 0 and 1.0. In addition, the number of hidden layers in

the networks was explored by performing hyperparameter searching across four different network architectures; networks with one, two, three and four hidden layers were tuned for both classification problems.

Five hundred replicate neural networks were generated using randomly initialized parameters with fifty thousand training epochs. Batches were randomly shuffled for each replicate based on their initialized batch size. After a complete set of five hundred replicate neural networks were trained, each replicate was evaluated based on its validation set accuracy. Parameters which greatly affected model performance had their parameter search ranges reduced and another set of five hundred replicates were generated for all four network architectures. For example, the learning rate was reduced from between 0 and 1.0 to between 0 and 0.1 after the first set of iterations for both the binary and multiclass neural network. The hyperparameter searching pipeline was completed three times for a total of one thousand five hundred replicates for both the binary and multiclass neural networks.

Once the optimal neural network architecture and parameters were identified for both the binary and multiclass networks, five hundred replicate optimal neural networks were trained for each classification problem. Each of the replicates used randomly split training, validation and test sets. This was performed to ensure that the seed used to split the data in the hyperparameter tuning pipeline was not producing outlier accuracies, and that the selected network parameters generated a robust network. Both optimal neural networks were then used to generate predictions on the proteins which were not used for training, the remaining predicted mitochondrial proteome, which can be found in *Supplementary File 5.3*. The list of novel predicted mitochondrial disease genes by both neural networks can be found summarised in *Appendix III*.

5.3. Results

5.3.1. Separation of the training sets using the input data array features

Hierarchical clustering of the training set proteins using their expression data can be seen in *Figure 5.1*. The positive training set proteins did not clearly cluster away from the negative set training proteins. However, the clade of proteins which showed a generally higher level of expression across all the tissues was almost entirely made up of the positive training set proteins. The clustering of the tissue types across the training set did not generate many meaningful clusters, the tree resulted in many small distinct clades which did not group together tissues of similar function. Two clades which did separate from the rest included the liver and kidney, and the skeletal muscle and heart muscle. All four of these tissues were found significantly different between the positive and negative training sets ($p < 0.001$).

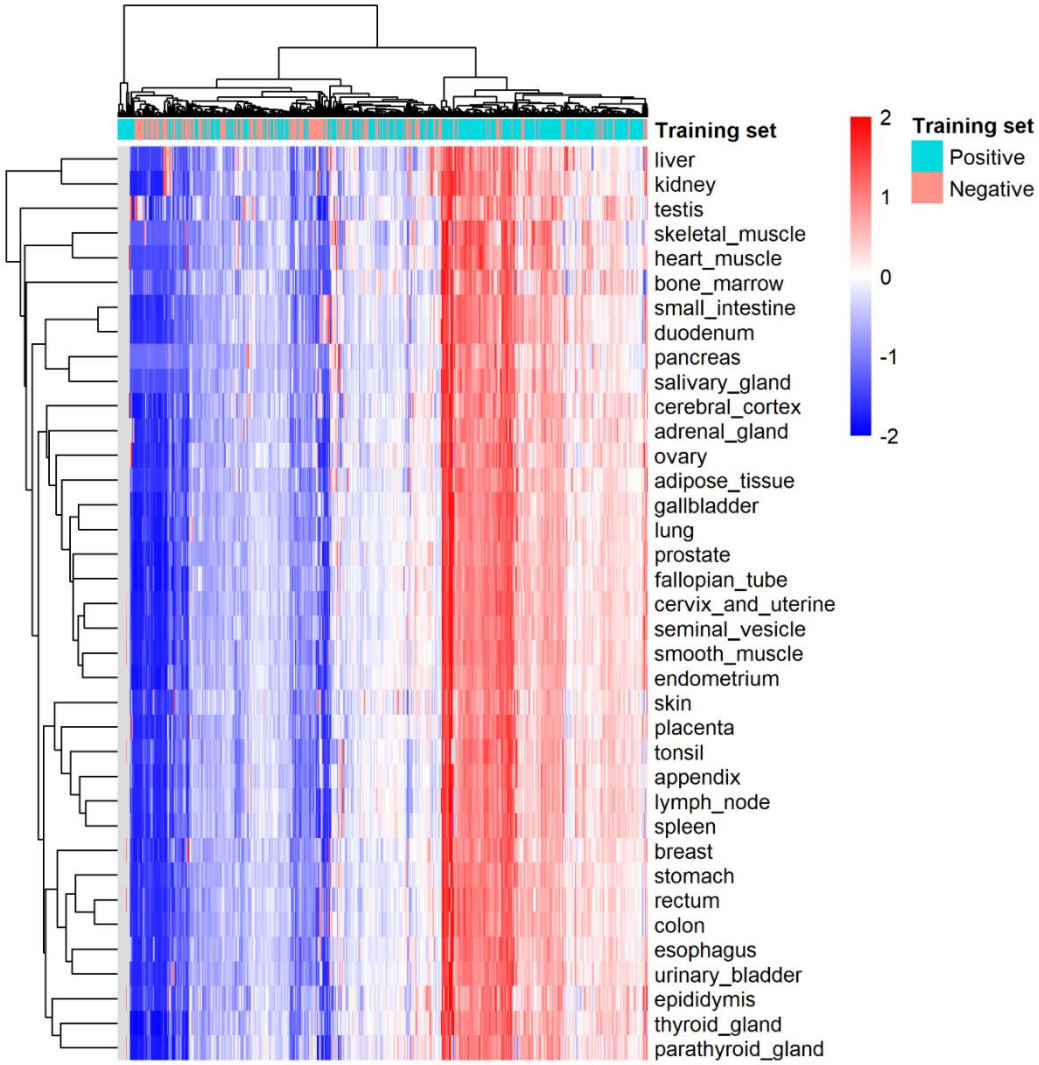


Figure 5.1. Hierarchical clustering of the training sets over the tissue expression features used in the input data array.

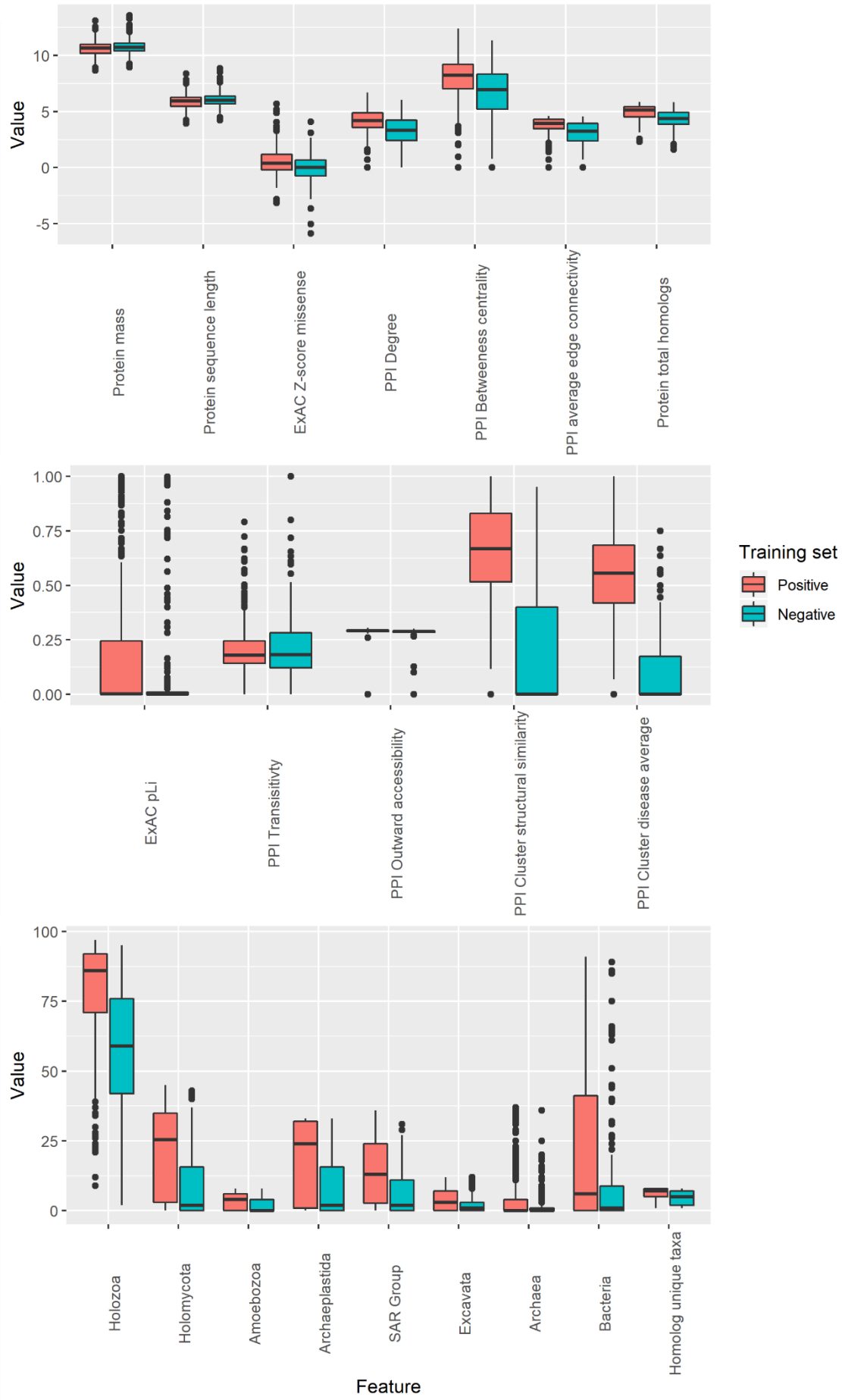


Figure 5.2. Boxplots of the continuous features used in the input data array over the training sets, excluding the tissue expression features.

Boxplots of the rest of the continuous features generated for the input data array from the protein-protein interaction network, evolutionary study and ExAC database can be seen in *Figure 5.2*. Hypothesis testing of each of the features identified only four features that were not significantly different ($p < 0.001$) between the positive and negative training sets. These were PPI transitivity ($p = 0.665$), the number of Archaea homologs ($p = 0.004$), the mass of the protein ($p = 0.008$) and the sequence length of the protein ($p = 0.01$). For the two ExAC database features, Z-score for missense and pLi, both found that disease genes had a higher value than the non-disease genes. For all the evolutionary study features, a similar relationship was found for the disease genes, they had a generally higher value than the non-disease genes. For the six other protein-protein interaction network generated features, degree, betweenness centrality, average edge connectivity, outward accessibility, cluster structural similarity average and cluster mitochondrial disease gene ratio average, the value for each of these was also found to be higher among disease genes compared to the non-disease genes.

The complete input array was used to train multiple replicate extreme gradient boosted decision trees. The importance score for each of the features was collected from each replicate and averaged. Two features had a much higher average score than the rest, cluster structural similarity average and cluster mitochondrial disease gene ratio average. Histograms showing the distribution of these two feature values across the positive and negative training set were generated (*Figure 5.3*). In addition, histograms showing the distribution of these two feature values across the proteins not used in either training set, the proteins which are not known to be disease or non-disease, were generated. The histograms clearly showed the difference in feature values between the negative and positive training sets. *Figure 5b* and *5d* include vertical lines on the distributions which highlight both the positive and negative training set averages. Both histograms highlighted the fact that there was only a small subset of the remaining mitochondrial proteome that had feature values close to and above the positive training set averages, the large majority had a value smaller than the negative set averages.

A second iteration of the pipeline for training the decision trees was performed after the two protein-protein interaction cluster features were removed to get an accurate

reflection of the importance scores of the remaining features, the results of which can be seen in *Figure 5.4*. The resulting average importance scores highlighted a wide variety of features as good for classification of the training sets. Three of the homolog features had a higher importance score than the rest; the homolog numbers of each protein in *Holozoa* and *Holomycota*, the two oldest taxa explored in the study, and the total number of identified homologs across all the taxa. Of the remaining protein-protein interaction network features transitivity, the only feature to have been found not significantly different, had a low importance average. The two ExAC features, Z-score for missense and pLi, had relatively high importance scores, as did protein sequence length. The tissue expression features had a huge number of tissues that had high importance score where spleen, lymph node, adrenal gland and cerebral cortex had particularly high values. The four tissues identified in the hierarchical clustering, liver, kidney, heart and skeletal muscle were all given lower importance scores, as were all the categorical variables.

T-SNE plots were generated using the complete input data array and predicted mitochondrial proteome. Proteins were labelled based on their training set assignments for both the binary and multiclass training sets (*Figure 5.5*). The positive and negative training sets used for the binary network showed good separation with many regions clearly belonging to only one training set. The t-SNE plot also identified multiple overlapping regions which were not unexpected as the input array was drastically reduced in dimensionality to only two dimensions. The recessive diseases made up most of the positive training set as seen in *Figure* 5b. All three of the disease types were spread among the t-SNE plot which suggested a high amount of intergroup variance, with the dominant diseases having the highest amount of variance despite only having twenty samples. The ‘other’ diseases had a small set of clusters of proteins whilst the recessive diseases were extremely similar to the positive set used to train the binary network, as expected due to the recessive diseases being the majority of the binary positive training set.

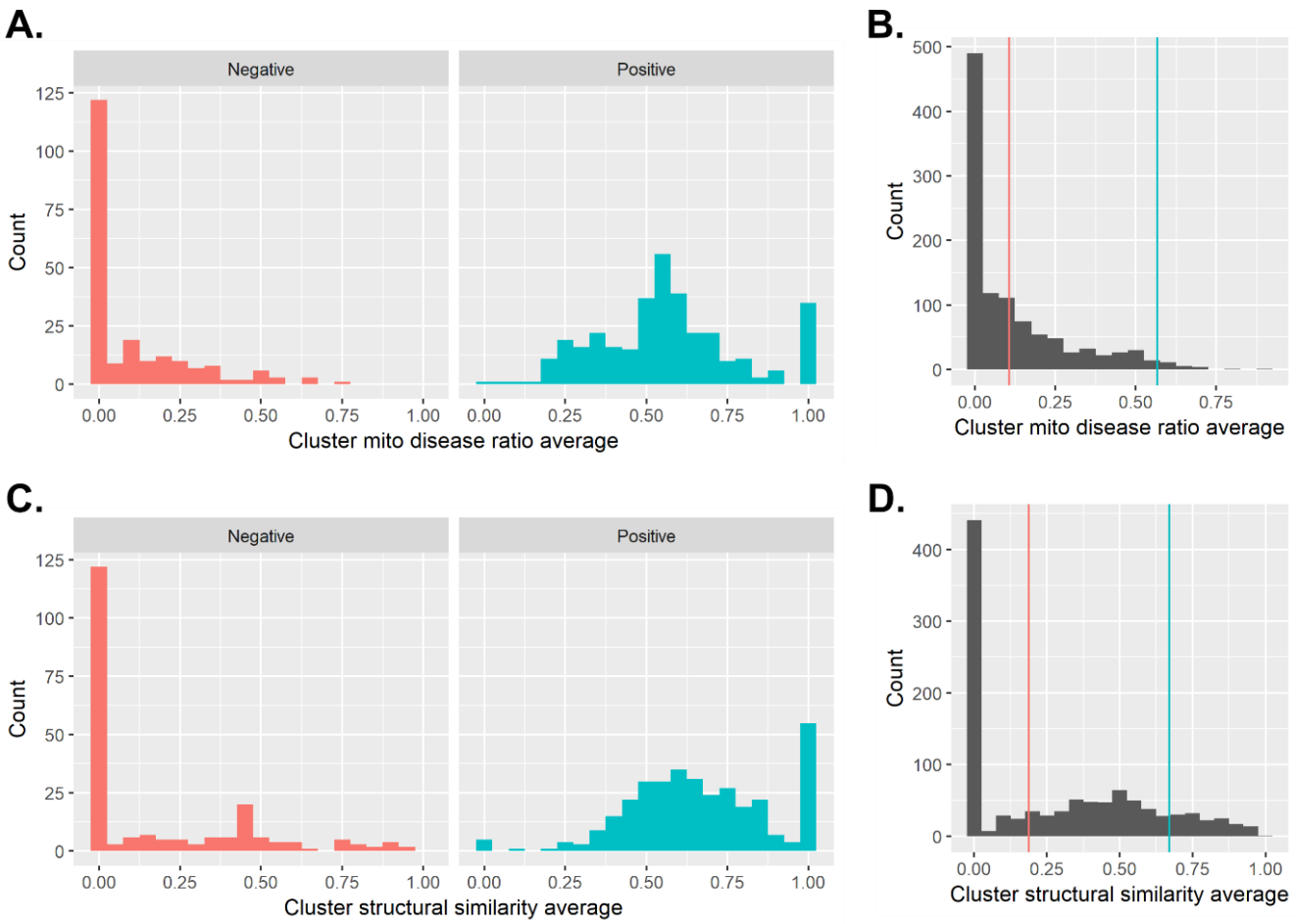


Figure 5.3. A) Histogram of the positive and negative training set cluster mitochondrial disease ratio average values **B)** Histogram of the remaining mitochondrial proteome cluster mitochondrial disease ratio average values, with lines indicating the negative (red) and positive (blue) training set averages. **C)** Histogram of the positive and negative training set cluster structural similarity average values **D)** Histogram of the remaining mitochondrial proteome cluster structural similarity average values, with lines indicating the negative (red) and positive (blue) training set averages.

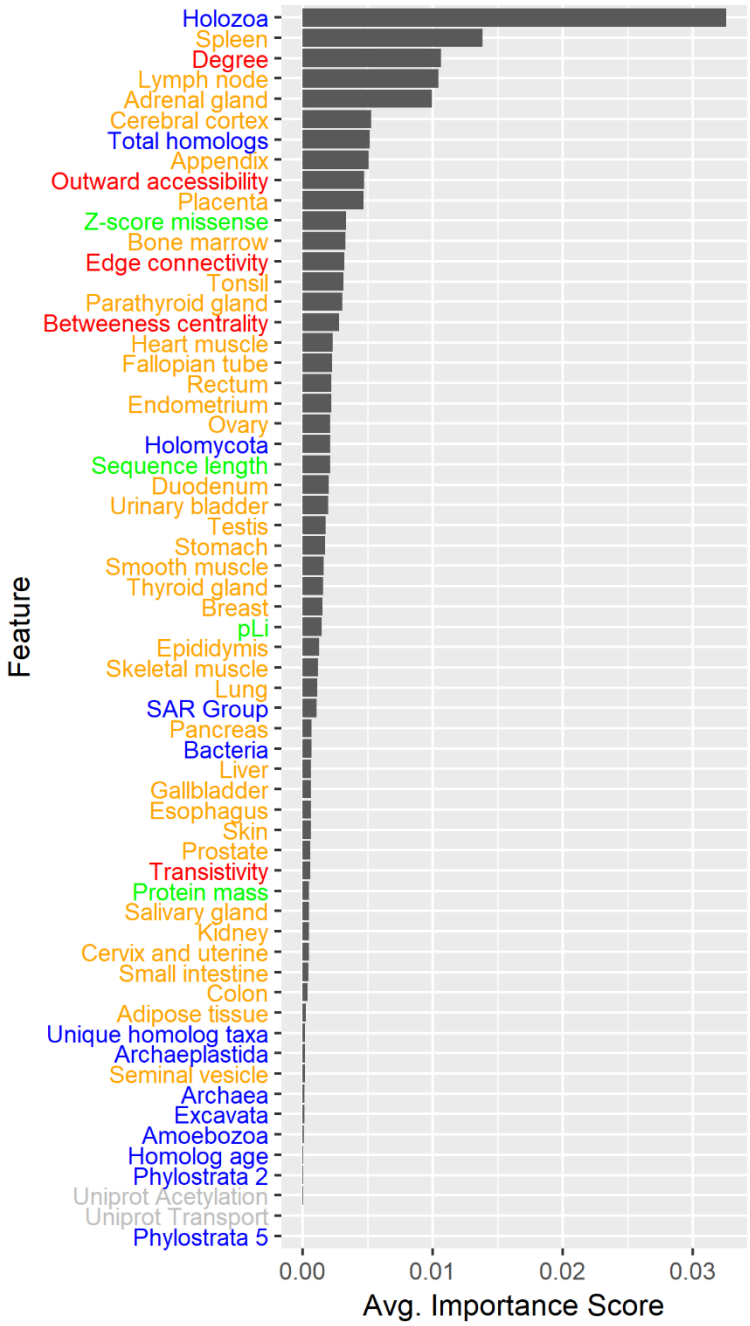


Figure 5.4. Average importance scores of the input array features when used to train extreme gradient boosted decision trees. Features are coloured by their means of collections; Red – PPI network, Yellow – Tissue expression, Blue – Evolutionary study, Green – Sequence information and ExAC, Grey – Annotations. 'Phylostrata' features are the one-hot representation of a proteins phylostrata of evolutionary origin.

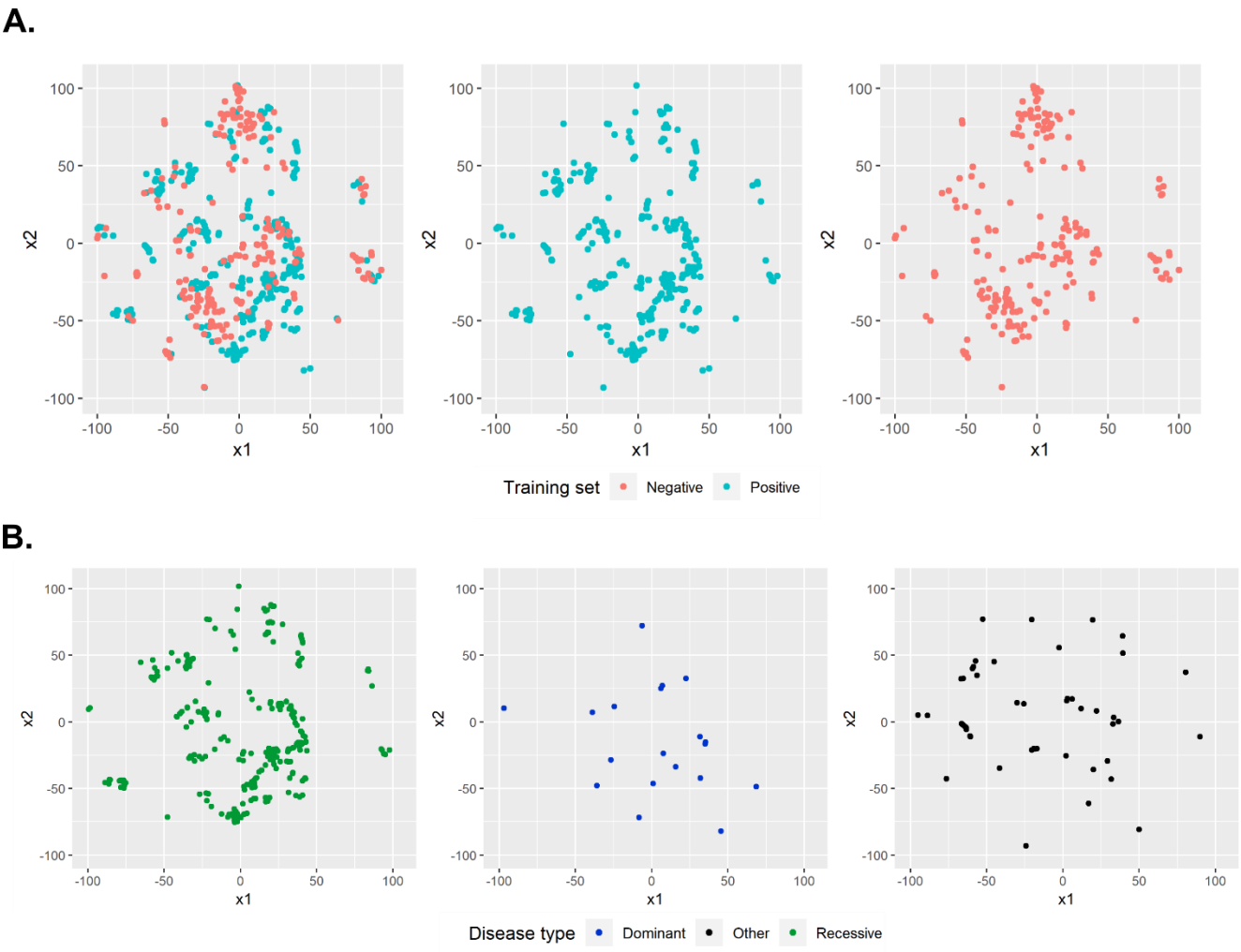


Figure 5.5. t-SNE plots of the positive and negative training sets using the complete input data array. **A)** Plots of the positive and negative training set used to train the binary neural network. **B)** Plots of the positive training set split into the three disease types which were used to train the multiclass neural network.

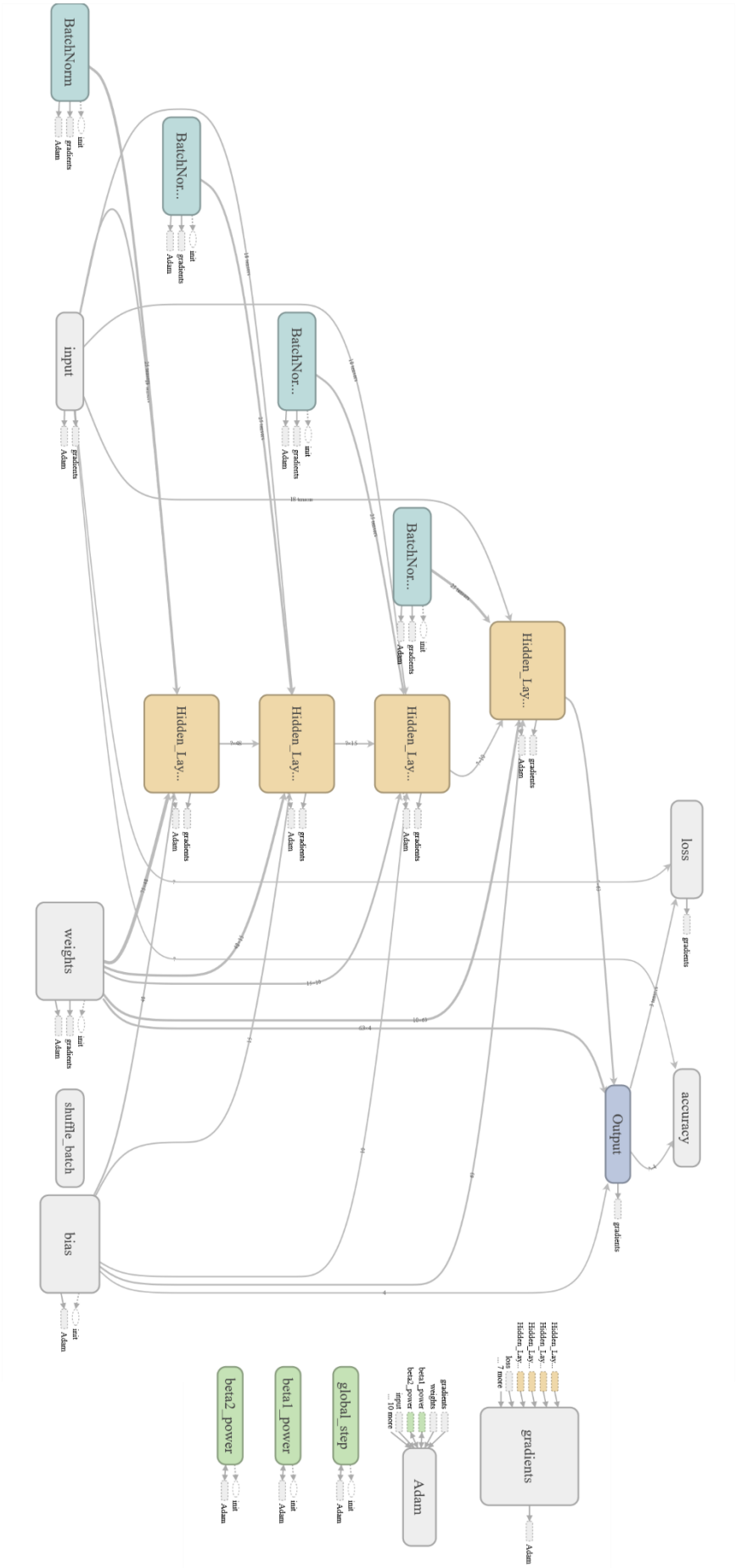


Figure 5.6. The network architecture of the multiclass neural network created in TensorFlow.

5.3.2. Training and validation of the neural networks after hyperparameter searching

Hyperparameter searching of the two neural networks identified the optimal hyperparameter values and network architectures for both the binary and multiclass classifications. *Figure 5.6* shows the network architecture for the multiclass network using four hidden layers, the binary neural network used a similar structure but with only two hidden layers. The final binary neural network was trained using the following optimal hyperparameters; learning rate = 0.026, batch size = 7, number of neurons in the first hidden layer = 35, number of neurons in the second hidden layer = 281, drop-out probability on the first hidden layer = 0.51 and drop-out probability on the second hidden layer = 0.85. The final multiclass neural network was trained using the following optimal hyperparameters; learning rate = 0.032, batch size = 21, number of neurons in the first hidden layer = 48, number of neurons in the second hidden layer = 15, number of neurons in the third hidden layer = 10, number of neurons in the fourth hidden layer = 63, drop-out probability on the first hidden layer = 0.76, drop-out probability on the second hidden layer = 0.60, drop-out probability on the third hidden layer = 0.48 and drop-out probability on the fourth hidden layer = 0.87.

The accuracies over the duration of training for the final optimal neural networks are shown in *Figure 5.7*. Throughout the duration of the training process, both neural networks showed a consistent improvement in both the training set and validation set accuracies. The final validation set accuracy for the binary neural network was 97.32% with a training accuracy of 95.51%. The validation loss of the binary network at the point of early stopping was 0.248, whilst the training loss was 0.133. The precision of the binary network was 0.953, the recall 0.936 and F1-score 0.944. The final validation set accuracy for the multiclass neural network was 82.14% with a training accuracy of 90.12%. The validation loss of the multiclass network was 0.935 whilst the training loss was 0.313. The precision and recall values for each of the classes in the multiclass network can be seen in *Table 5.2*. The average precision value for the network was 0.639 and average recall 0.580 giving an F1-score of 0.608. The generalization accuracy of the final binary neural network was 90.18%, whilst the multiclass neural network was 81.25%

The breakdown of the disease class values for the positive training set proteins and the non-disease class value for the negative training set proteins assigned by the final binary neural network are shown in *Figure 5.8*. The binary network correctly classified 322 of the positive training set as disease genes, with 288 of those being assigned a value greater than or equal to 0.75, leaving 22 incorrectly classified as non-disease (*Table 5.3*). For the negative training set, 198 proteins were correctly assigned with 171 of these having a high class score (>= 0.75), leaving 16 proteins that were incorrectly classified. The breakdown of the final multiclass assignment scores for each of the training proteins belonging to the four classes are shown in *Figure 5.9*, each training protein has only been shown with their respective class value i.e. recessive training proteins are only shown in the recessive class score histogram. The multiclass network correctly assigned 207 of the non-disease proteins with 203 of these having a non-disease value greater than or equal to 0.5, leaving 7 incorrectly classified which were all assigned as recessive disease (*Table 5.4*). The recessive training set had 256 correctly identified, leaving 17 incorrect, with 246 of these having a high recessive class value (>= 0.5). The incorrectly classified proteins had 15 assigned as non-disease whilst two were assigned as other disease. In the dominant disease training set, none of the proteins were correctly classified as dominant. Instead, 15 of the dominant diseases were assigned as recessive, two as non-disease and 3 as other disease. The other disease set contained 21 correctly classified proteins with 18 of these having a high other disease class value (>= 0.5). The 30 incorrectly classified other disease genes had 22 assigned as recessive, 7 as non-disease and one as dominant.

The accuracies of the fully trained replicate network generated using the optimal hyperparameters and network architecture for both the binary and multiclass networks are shown in *Figure 5.10*. The binary neural network achieved an average validation accuracy of 93.45% with an average training accuracy of 90.10%. The multiclass neural network achieved an average validation accuracy of 83.99% with an average training accuracy of 86.43%.

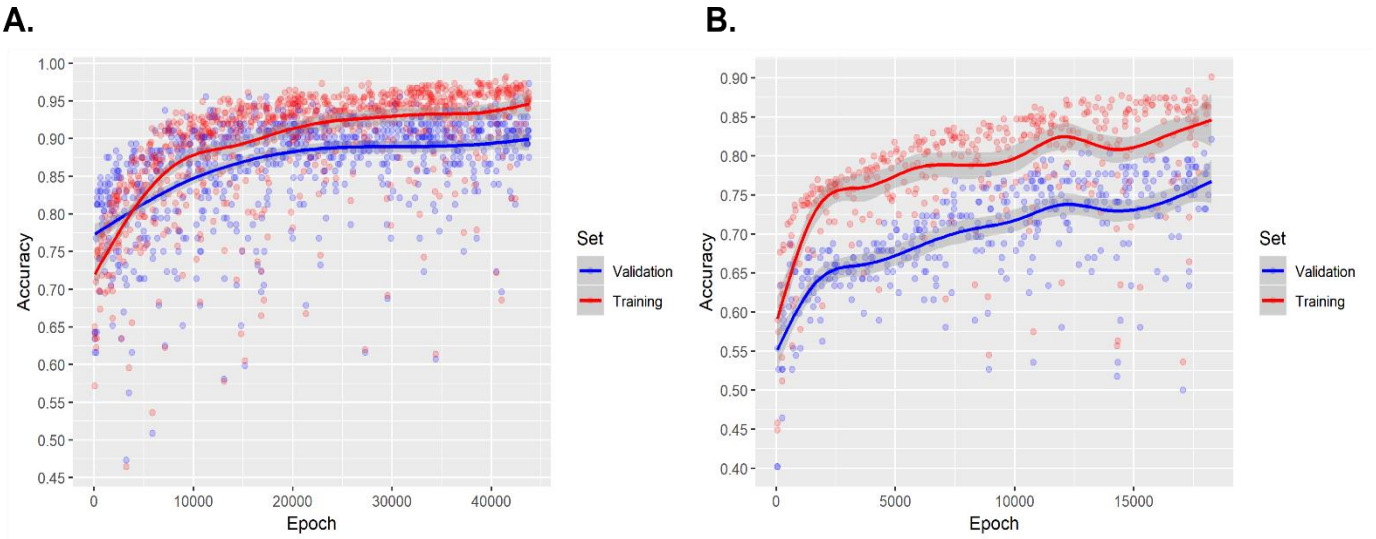


Figure 5.7. Training of both the binary and multiclass neural networks over the duration of the training epochs. **A)** Binary neural network. **B)** Multiclass neural network

	Non-disease	Recessive	Dominant	Other
Precision	0.896	0.853	0	0.808
Recall	0.967	0.938	0	0.412

Table 5.2. The precision and recall values of the multiclass neural network for each of the classes.

		Predicted	
		Disease	**Non Disease**
Actual	**Disease**	322	22
	Non Disease	16	198

Table 5.3. Truth table for the training sets predicted using the binary neural network

		Predicted			
		Non disease	**Recessive**	**Dominant**	**Other**
Actual	**Non disease**	207	7	0	0
	Recessive	15	256	0	2
	Dominant	2	15	0	3
	Other	7	22	1	21

Table 5.4. Truth table for the training sets predicted using the multiclass neural network

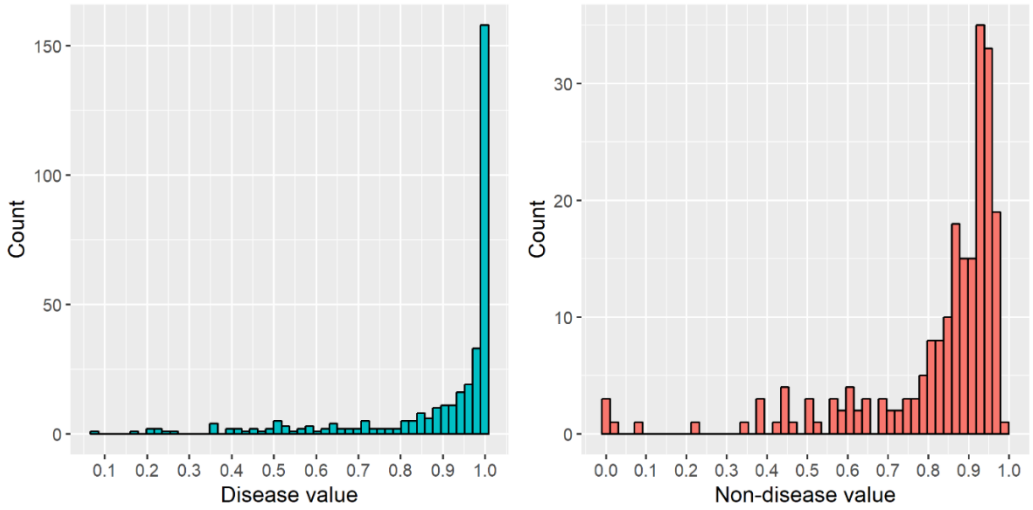


Figure 5.8. Histograms of disease class values assigned to the positive training set and non-disease class values assigned to the negative training set by the binary network.

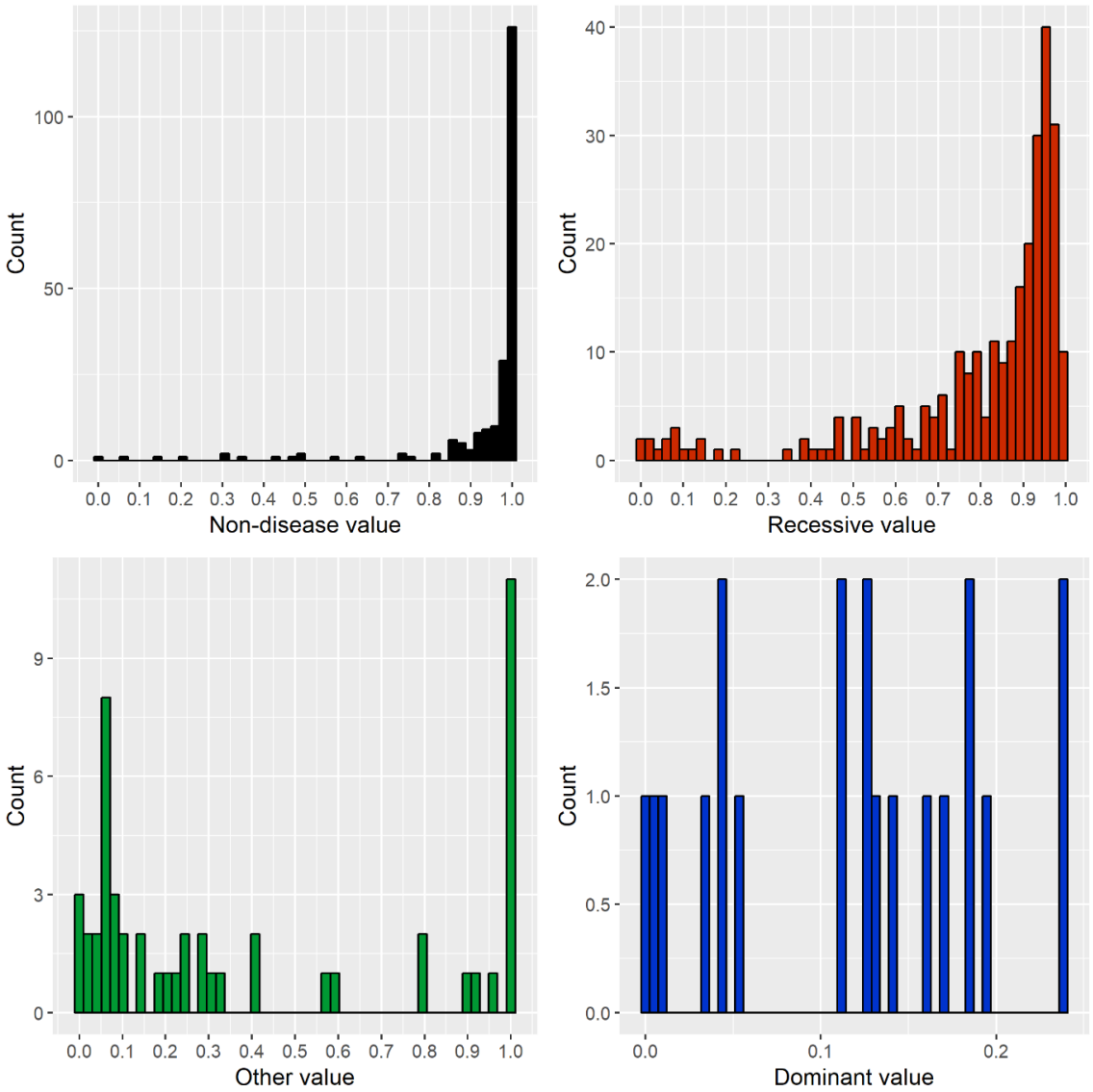


Figure 5.9. Histograms of the four class values assigned to each of the proteins belonging to each of the specific class training proteins using the multiclass neural network.

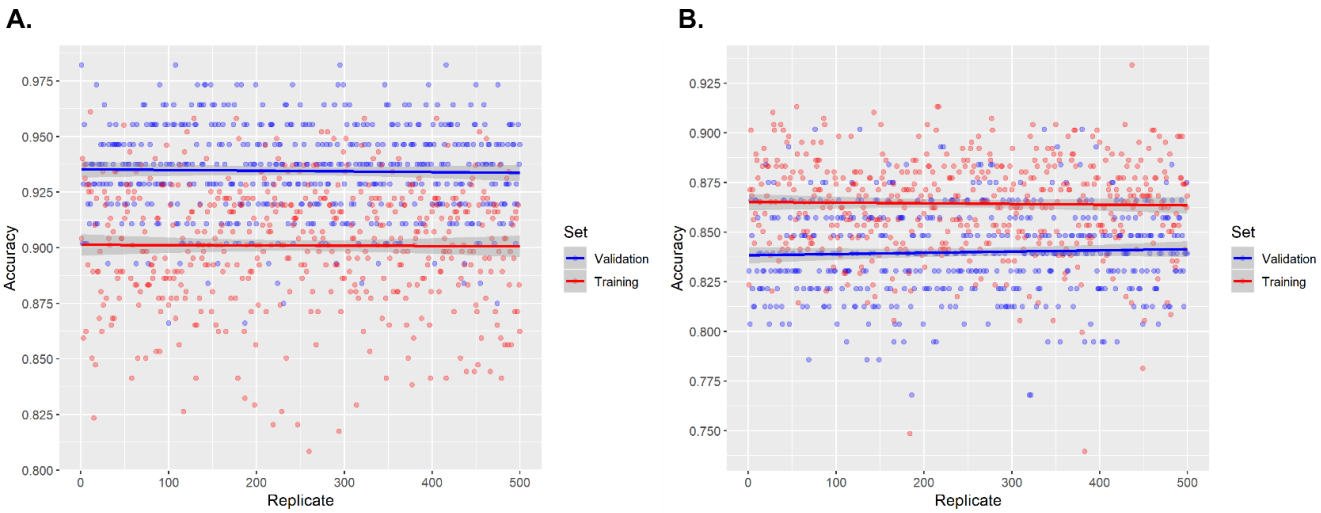


Figure 5.10. Training and validation set accuracies of 500 replicate optimal neural networks generated using random training set splits **A)** Binary neural network. **B)** Multiclass neural network.

5.3.3. Predictions on the remaining predicted mitochondrial proteome using the trained neural networks

The remaining 1,067 proteins belonging to the predicted mitochondrial proteome which were not used in any of the training sets were fed through both the binary and multiclass neural networks to generate predictions to identify novel disease causing. The final binary neural network assigned 422 as disease genes with 330 of these having a disease class score greater than or equal to 0.75, and 184 having a score greater than or equal to 0.99 (*Figure 5.11*). The remaining 645 proteins were assigned to the non-disease class, with only 82 of these having a disease class score greater than or equal to 0.3 and only 5 having a disease score greater than 0.5.

The multiclass neural network assigned 733 proteins as non-disease, with 722 of these having a non-disease class value of greater than or equal to 0.5 and 683 having a score greater than or equal to 0.75. The remaining proteins were assigned as follows; 232 recessive disease, 92 other disease and 10 dominant disease. The set of recessive genes contained 191 genes assigned a recessive class score greater than or equal to 0.5 with 113 of those having a score greater than or equal to 0.75. The set of ‘other’ disease genes contained 71 genes assigned an ‘other’ disease class score greater than or equal to 0.5 with 38 of those having a score

greater than or equal to 0.9. The predicted dominant disease genes contained only two genes with a dominant class score greater than or equal to 0.5.

By summing together the class assignment scores of the three disease types used in the multiclass neural network, a disease class score was calculated for each protein based on the multiclass neural network predictions. The disease class value for each protein assigned by both the binary and multiclass networks were plotted on the t-SNE plot generated using the compete input array (*Figure 5.13*). The disease scores from both networks show a similar distribution around the t-SNE plot, showing a similarity in their predicted disease vs. predicted non-disease assignments. The two neural networks agree upon 304 proteins being predicted as disease-causing, and 615 proteins as non-disease causing (*Figure 5.14*). The binary network assigned 118 proteins as disease-causing which were assigned as non-disease in the multiclass network, whilst the multiclass network assigned only 30 proteins as disease-causing which were assigned as non-disease by the binary neural network.

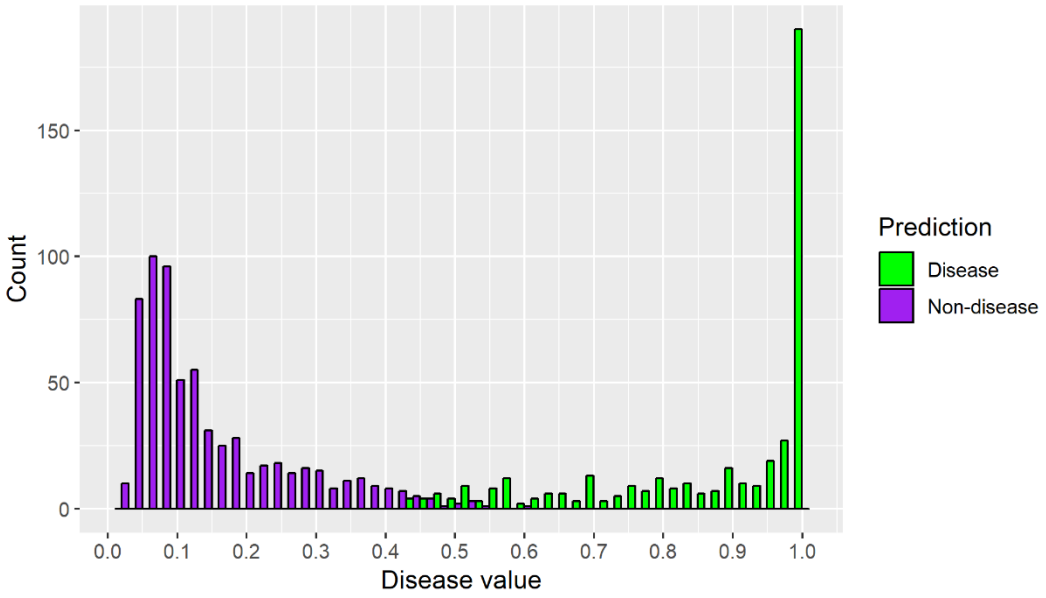


Figure 5.11. Histogram of the disease class values assigned to the remaining mitochondrial proteome using the binary neural network. Those predicted as disease-causing are green, with the predicted non-disease in purple.

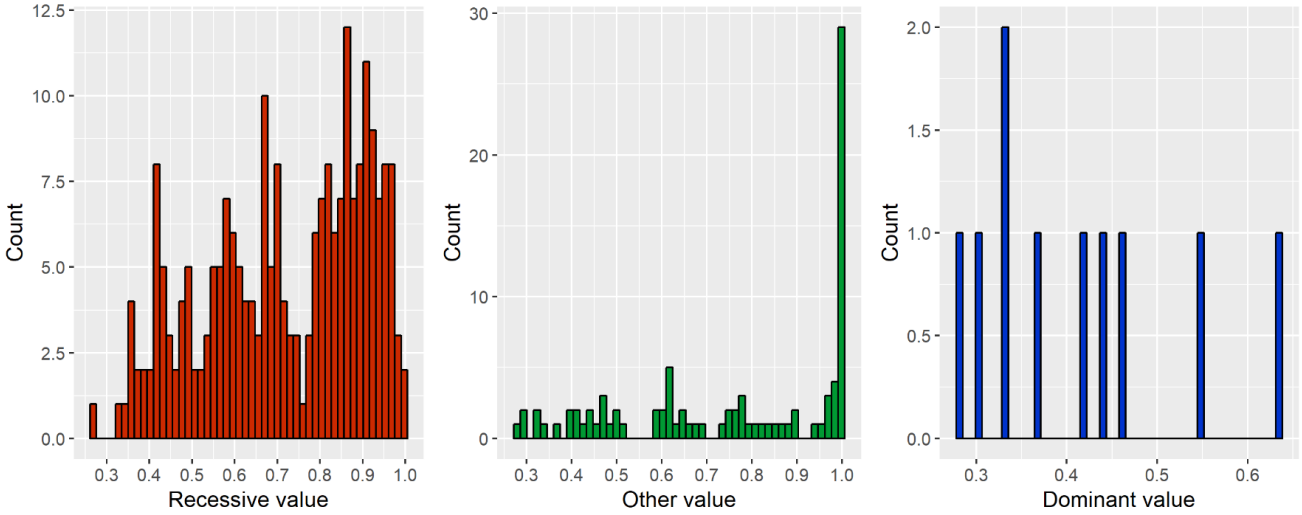


Figure 5.12. Histograms of the disease class values assigned to the remaining mitochondrial proteome using the multiclass neural network. Only proteins which were assigned to each of the three classes are shown in their respective histogram.

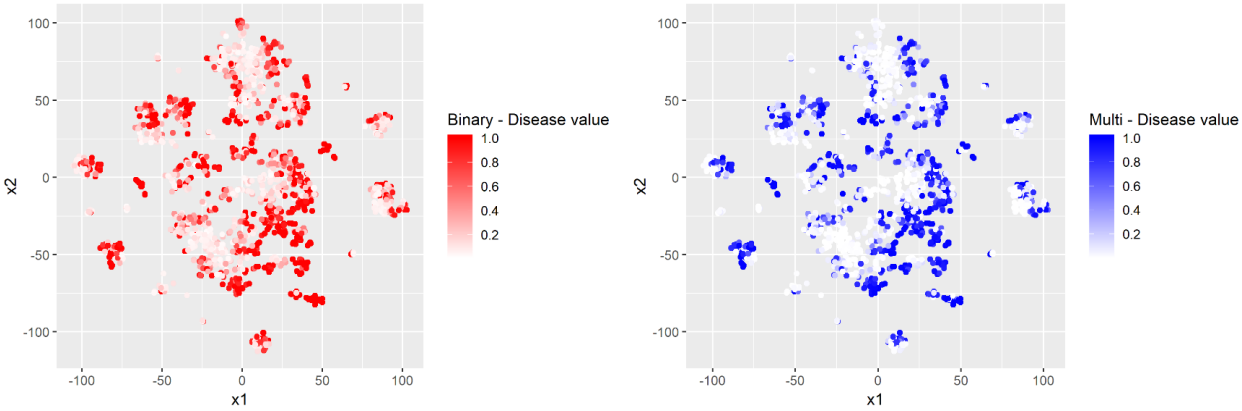


Figure 5.13. The predicted disease values for the remaining mitochondrial proteome from both the multiclass and binary neural networks overlaid onto the input data array t-SNE plot.

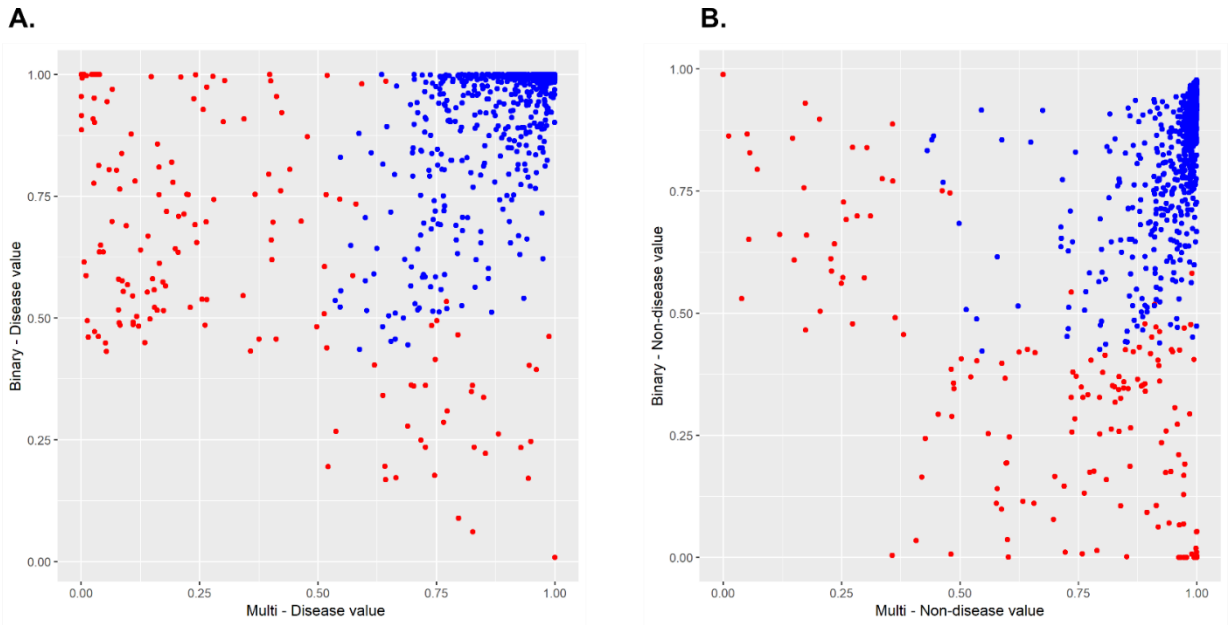


Figure 5.14. Scatter plot of the predicted disease and non-disease values of the binary and multiclass neural networks. In blue are the proteins that are predicted as the same class by both networks, whilst red are predicted by just one of the networks. **A)** Predicted disease genes. **B)** Predicted non-disease genes.

5.4. Discussion

5.4.1. Properties of mitochondrial disease genes

Hierarchical clustering of the tissue expression features across the entire training set shows that using tissue expression data alone is not enough to clearly differentiate the known mitochondrial disease genes from the negative training set (*Figure 5.1*). The result is not unexpected as the set of known mitochondrial disease genes contains many different functional types of protein which are known to have consistently different expression profiles across all tissues in humans. For example, the known disease genes contain ribosomes, which are known to have consistently high expression in humans, and transporters, which have a consistently lower expression in humans. However, the right clade of the clustered training proteins does indicate that highly expressed proteins may have a higher propensity to be disease genes. Highly expressed proteins have been shown to evolve at a much slower rate than those which have lower expression [309]. Slower evolution of a protein leads to greater conservation of the protein among species akin to a form of evolutionary pressure usually put on important proteins. The importance of these genes thus identifies a potential explanation for why the highly expressed proteins have an increased likelihood of being disease-causing genes.

When clustering tissue specific expression across any set of human genes, it would be expected that tissues of a similar functional role, such as the set of hormone producing tissues, would cluster together. The hierarchical clustering of the tissues across the training set did not follow this pattern in most cases, the tree consisted of mostly small, disorganised clades. This highlights that in most tissues, both the disease and non-disease genes have highly variable tissue specific expression profiles. The exception to this can be seen by the clustering of the liver and kidney, and the skeletal muscle and heart, into distinct clades separated from the rest of the tree. These four tissues are some of the most energetically demanding tissues in humans and facilitate many integral metabolic functions such as producing urea in the liver and kidneys. These tissues are also the most affected tissues during mitochondrial disease and are regularly studied when investigating mitochondrial dysfunction. This clustering behaviour, and the fact that all four of these tissues were identified as significantly different between the positive and negative training sets,

identifies these four tissues as having the potential to predict mitochondrial disease genes. Since the disease genes showed a consistently higher expression across all the tissues using hierarchical clustering, which was a relatively simple method for separating the training sets, the complete set of tissue expression features were used in the training of both neural networks.

Four of the continuous features generated were not found to be significantly different between the two training sets using a very conservative p-value threshold of p = 0.001. However, if a more lenient threshold value were selected such as p = 0.05, only one of these would remain not significant, PPI transitivity. The near significant difference of protein sequence length and mass identifies that the known disease proteins are on average much smaller proteins than the negative training proteins (*Figure 5.2*). In a study in yeast, highly expressed proteins were identified to be generally smaller proteins [310]. This identifies a potential link between disease genes having generally higher expression and being smaller. The pLi ExAC database feature identified that the known disease genes have on average a higher pLi score which was expected as the sole purpose of the metric is to identify genes which are intolerant to loss-of-function variants. However, the large variance seen in both the training sets shows that this value is not entirely dependable as many of the known disease genes had low pLi values, whilst many of the negative training set had high pLi values. The known disease genes also had higher Z-score for missense than the negative training set. This meant that based on the ExAC population study, the disease genes had fewer observed variants than would be expected in the healthy population i.e. the disease proteins had a higher intolerance to variation. This suggests that the disease genes have undergone a high amount of evolutionary pressure and are therefore proteins of importance that, once disrupted, lead to disease.

The complete set of protein-protein interaction features, including the one which was not significantly different, identified many network features of the known mitochondrial disease genes which mostly agree with theorems on the entire set of known human disease genes. The degree feature showed that the known mitochondrial disease genes had on average more interactors than the negative training proteins, a theorem which had previously been shown for all known human

disease genes [275]. The set of known mitochondrial disease genes generally clustered together, highlighted by the cluster disease ratio average feature, and within the clusters the disease genes on average shared more similar interactors with other known disease genes than the negative training set, shown by the cluster structural similarity average. Both theorems are consistent among all known human disease genes [275, 272]. In addition, the average edge connectivity feature suggests that the mitochondrial disease genes are generally more highly connected to the other known disease genes than the negative training set. The positional importance of the known mitochondrial disease genes is shown by the betweenness centrality and outward accessibility feature. The mitochondrial disease genes had on average higher betweenness centrality and outward accessibility which identifies that the known disease proteins are placed in highly positionally important locations within the network (e.g. as the only protein which connects two clusters generating a choke point in which many pathways must go through), a theorem which was proven for all known human disease genes [272]. The only contradiction to current theorems of all known human disease genes comes from the lack of significant difference in the transitivity feature. Human disease genes have been shown on average to avoid highly connected areas [272]. The highly variable and non-significant difference in the transitivity between the positive and negative training sets identified that this is not necessarily true for the known mitochondrial disease genes.

The features generated using the evolutionary study showed the exact same results reported in the study [279]. The known mitochondrial disease genes had on average a much higher total number of homologs than the negative training set. In addition, the homologs of the known disease genes were identified among more unique taxa than the homologs of the negative training set. The known disease genes had significantly more homologs than the negative training set across all the taxa studied, but the difference was much larger in the older taxa. In general, these results suggest that the known mitochondrial disease genes originated much earlier in evolutionary history and have been conserved across eukaryote evolution at a greater rate than the negative training set proteins. This result implies a higher importance of the known mitochondrial disease genes than the negative training set.

Based on training set significant differences, almost all the continuous features could be used individually to predict a genes' disease status. However, using only one of the features for predictions would vastly oversimplify the problem of predicting mitochondrial disease. The use of a complex algorithm such as a neural network allows for the identification of complex relationships between all the features for predicting disease and would lead to a much high confidence in the predictions. The linear separability of the training sets using each of the continuous features highlights each features viability for use in training a neural network for classification of disease genes. Therefore, the complete set of continuous features were used in the training of both neural networks. Despite being not significantly different between the positive and negative training sets, the transitivity feature was included in the training process as hypothesis testing is only a simple means for identifying separability and does not disqualify the feature from containing useful information when used in conjunction with other features within a more complex algorithm for classification.

5.4.2. Viability of the input data array for training the neural networks

The first iteration of gradient boosted decision trees identified the two protein-protein interaction cluster features as the two best features for classification of the positive and negative training sets by a huge margin. As the gap in average importance score between these two features and the rest was so great, the features needed to be investigated further to see if the high classification power assigned by the decision trees was justified in both a machine learning and biological viewpoint. If the assignment was entirely justified, it needs to be ensured that the two features would not have a huge effect on the training and final predictions of the neural network as the purpose of the network is to identify complex relationships between the complete set of input features and not to overfit to two features.

The distribution of the two feature values across the positive and negative training sets clearly justified why the two features were assigned a high classification power by the decision trees from a machine learning viewpoint (*Figure 5.3*). In the complete clustered protein-protein interaction network which was used to calculate the two

features, the average cluster size of any cluster which had at least one mitochondrial protein was 25, each of which contained on average 4 mitochondrial proteins. The average ‘cluster mito disease ratio average’ value for the negative training set was 0.11, whilst the positive training set was 0.57. Based on the average of 4 mitochondrial genes per cluster, the known mitochondrial disease genes therefore clustered with 2.2 other known disease genes on average, whilst the negative training proteins only clustered with 0.4 known disease genes on average. Therefore, the known mitochondrial disease genes clustered with a substantially larger number of known disease genes than the negative training proteins, which rarely clustered with a single known disease gene. Given a protein of unknown disease association, if the protein were to cluster with a large proportion of known mitochondrial disease genes, from a biological viewpoint it would not be unreasonable to assign the protein a high probability of being disease-causing.

The ‘average cluster structural similarity average’ for the negative training set was 0.19, whilst the positive training set was 0.67. This means that the known mitochondrial disease genes share 67% of the same interactors on average with all other known disease genes within their clusters, whilst the negative training proteins only shared 19% of the same interactors as the known disease genes within their clusters. This indicated that within the clusters, the known mitochondrial disease genes were positionally close to each other, whilst any negative training set protein which clustered with at least one known disease gene was positionally distant from the disease gene(s). Given a protein of unknown disease association, if the protein were to cluster with at least one known disease gene and shared many of the same interactors within the cluster it would be biologically reasonable to assign a high association for disease to the protein. These arguments show that the assigned high importance score to these two features by the decision trees was entirely justified and would be expected in the neural network.

The variance in the distribution of the two feature values for the training sets was reasonably large, particularly for the positive training sets. In combination with the regularisation methods used in the training process of the neural network, this variance would help prevent the neural network from overfitting to these two features. In addition, the distribution of the remaining predicted mitochondrial

proteome, the set of proteins which were ultimately predicted on, showed that there was only a very small number of proteins which had feature values anywhere near the positive training set averages. The high classification power of the features was therefore not expected to cause a drastic effect on the final predictions. Both features were therefore left in the input data array unchanged and used in the training of the neural networks.

The second iteration of trained extreme gradient boosted decision trees gave a large set of features with comparatively high average importance score (*Figure 5.4*). The top scoring set of features contained features collected from many of the different data sources which gave support for each of their inclusion in the neural network input data array. Tissue expression features were highly favoured among the data array, making up six of the top ten features. The tissues which ranked highly did not have any obvious links to mitochondrial function and showed no common trend in function. Three of the four tissues which stood out in the hierarchical clustering which are affected by mitochondrial disease the most, the liver, kidney and skeletal muscle, were all assigned low importance scores whilst the heart muscle ranked in the middle. This gave support for the inclusion of the complete set of tissue expression data as the decision trees identified relationships in unexpected tissues that could be used to classify the two training sets. The top ranking feature was the number of homologs found in *Holozoa* collected from the evolutionary study. Two other features from the study also ranked relatively highly, the total number of homologs and the number of homologs in *Holomycota*. *Holozoa* and *Holomycota* were the two oldest taxa examined in the evolutionary study which further highlights the theory that mitochondrial disease genes are much older and more conserved than the negative training set genes. The only remaining protein-protein interaction feature which was assigned a low importance score was the only non-significant feature, protein transitivity, which made the result unsurprising. The entire set of categorical variables were included in the decision tree training array, but they were all assigned extremely low importance scores. This was most likely due to the decision tree method rather than a reflection of their classification power as decision trees are highly prone to overfitting, causing the algorithm to prefer the continuous variables rather than the categorical.

The t-SNE plot of the input data array and training set proteins showed good separation of the positive and negative training sets (*Figure 5.5*). The plot contained distinct regions for each of the training sets which showed that the input data array contained enough valid information to be able to classify mitochondrial disease genes with a good degree of accuracy. The plot did contain regions where the two training sets overlapped which identified that a subset of the positive and negative training sets were similar across the input data array and highlights the complexity of the classification problem. The t-SNE plot of the separated positive training set into the three disease classes highlights the high intergroup variance in all three of the classes. The high variance of the dominant diseases and the large overlap in the positions of the dominant disease genes with the recessive disease genes reflects the difficulty in differentiating the dominant from recessive disease genes. Compounded by the fact that there is currently only a small set of known dominant mitochondrial disease genes, these factors indicate that the training of the multiclass neural network for predicting dominant diseases would most likely be extremely difficult.

Overall, each feature showed good separation of the two training sets based on hypothesis testing and the combined input data array separated the training sets well on the t-SNE plot. The complex nature of the problem and the combination of linearly separable features was the perfect scenario for use in a complex machine learning algorithm, such as a feed-forward neural network. Therefore, the input data array and training sets were used to train the binary neural network, predicting disease or non-disease, and the multiclass neural network, predicting non-disease, recessive, dominant or ‘other’ disease. The added complexity of the separated positive training set meant that the multiclass neural network was expected to perform badly on the dominant disease genes but was explored due to the lack of current means to diagnose dominant diseases and the importance to do so.

5.4.3. Machine learning success of the final trained neural networks

The end goal of both neural networks was to generate disease association scores for each protein in the predicted mitochondrial proteome, giving a means to filter through

a patient's list of variants to find the disease-causing variant. The miss-classification of a disease gene as non-disease would have severe consequences and would lead to a patient receiving no diagnosis. This factor led to the decision to use the accuracy of the neural networks as the metric for evaluating network performance, instead of the loss value. Early stopping was employed during the final training of the neural networks using accuracy as the network performance metric to ensure that the networks reached their maximum validation accuracy without excessive training epochs to reduce the chance of overfitting. The changes in the training set accuracy and validation set accuracy over the duration of the training epochs prior to early stopping were also investigated for any signs of overfitting.

Both neural networks showed a consistent increase in the training and validation set accuracies across their complete set of training epochs (*Figure 5.7*). In addition, the final validation and training set accuracies showed only a small difference of around 5% for both neural networks. Similarly, the final validation and training set loss had only a small difference at the point of early stopping for both networks. The multiclass neural network performed worse than the binary network as expected, with the training accuracy constantly higher than the validation accuracy during training and at the point of early stopping, a potential sign of underfitting. The multiclass neural network showing signs of underfitting was an expected result as the dominant disease genes had such a large ingroup variance and small training set size. The average accuracies of the replicate neural networks are similar to the recorded final neural network accuracies (*Figure 5.10*) which highlights the robustness of the neural network hyperparameters and architecture to the randomised training set splits. This combined set of results suggests that all the methods used to regularize the neural network, such as drop-out and batch normalization, were working correctly and that neither network had any obvious overfitting.

The miss-classification of a known disease gene as non-disease is a greater problem when diagnosing a patient than the reverse situation. This identifies that a high recall is the most importance metric for both neural networks. The binary neural network achieved a high value for all the network performance metrics along with a high generalization accuracy. The network showed no sign of overfitting and was

therefore successfully trained for classification of mitochondrial disease genes. The network assigned very high (≥0.75) class values to 89% of the correctly assigned known mitochondrial disease genes (*Figure 5.8*), and 86% of the correctly assigned negative training set proteins. Therefore, the binary neural network had extremely high confidence in the vast majority of its class assignments which gives confidence in the predicted class values, particularly for those assigned a high value.

The multiclass neural network had mixed success across the groups based on the network performance metrics but managed to attain a reasonable generalization accuracy (*Table 5.2*). The negative training set and recessive disease training set achieved good precision and high recall values. The overall network performance metrics were dragged down by the network's performance on the dominant and 'other' disease training set. The domain disease training set was particularly bad with not even a single gene correctly predicted as dominant, a result which was expected as the training set was small and had a large amount of ingroup variance. The multiclass network showed a high preference towards predicting all the known disease genes as recessive diseases which caused the recall of the 'other' and dominant disease training sets to be low. Around 75% of the incorrectly classified dominant and 'other' disease training set genes were classified as recessive diseases. The propensity of the multiclass network to classify all known disease genes as recessive was caused by the imbalanced disease training set sizes as recessive diseases make up 79% of all known mitochondrial diseases.

However, 14 of the genes which belonged to the 'other' disease training set and one gene which belonged to the dominant disease set which were incorrectly classified as recessive disease but are in fact X-linked recessive diseases or diseases which were found as both dominant and recessive. These predictions by the network are therefore technically not incorrect and highlight an improvement which could be made to the multiclass network. The 'other' disease group contained genes which do not clearly fall into the recessive or dominant disease group and were therefore separated into their own category as it was unknown whether these genes were vastly different across the input data array. The separation of the 'other' disease group into the recessive and dominant disease group would potentially lead to a better performing neural network for classification of disease type as it would reduce

the number of trained classes and increase the size of the dominant disease training set, although the dominant disease training set would still be small and most likely suffer from similar underfitting problems.

The multiclass neural network assigned 98% of the correctly classified negative, 96% of the correctly classified recessive and 85% of the correctly classified other training set proteins a high class score (≥0.5). The network therefore had a high confidence in the class assignment for these three classes (*Figure 5.9*). Combined with the high precision and recall values for the negative training set and recessive diseases suggests that the predictions made by the network for these two classes have some validity. In fact, the predictions of the multiclass neural network when differentiating the known disease genes from the negative training set showed similar accuracies to the binary neural network. By combining all three disease classes scores to create a single multiclass disease class score and using the negative training set, the multiclass neural network achieved a precision of 0.979 and recall of 0.930 when classifying the known mitochondrial disease genes and the negative training set. The multiclass neural network can therefore be used as a second reference for predicting gene disease association with the added ability of identifying a gene's similarity to the know recessive genes, offering a potential way to identify new dominant disease genes.

5.4.4. Prediction results on the predicted mitochondrial proteome

The set of known mitochondrial disease genes assigned the highest disease class score for both neural networks are shown in *Table 5.5*, using the sum of the three multiclass disease class scores as the multiclass overall disease score. Both sets of genes were extremely similar, containing only the mitochondrial OXPHOS complex subunits. When patients are initially screened, these are the exact set of genes that are always investigated first and any identified mutation in an OXPHOS complex subunit gene is immediately suspected as the disease causing variant. Therefore, the identification of these disease genes as the highest disease causing genes out of the set of 344 known mitochondrial disease genes was an accurate representation of our current understanding of mitochondrial diseases and gave confidence to the

predictions of disease association by both networks on the remaining predicted mitochondrial proteome.

Binary	Multiclass
MT-ATP6	MT-ATP6
MT-CO1	MT-ATP8
MT-CO2	MT-CO1
MT-CO3	MT-CO2
MT-CYB	MT-CO3
MT-ND2	MT-CYB
MT-ND3	MT-ND2
MT-ND4L	MT-ND3
MT-ND5	MT-ND4L
MT-ND6	MT-ND5

Table 5.5. The top ten disease class value genes by both networks which are known disease genes.

The binary network predicted 422 disease associated genes out of the remaining 1,067 mitochondrial genes, 44% of these were assigned extremely high disease class scores similar to the scores assigned to the known disease causing mitochondrial complex subunits (*Figure 5.11*). Within this set of extremely high scoring disease associated genes was the set of remaining mitochondrial complex subunits which currently have no disease association, such as MT-ND1 and COX6C. These subunits would also be used in early screening during the patient diagnosis process along with the already known disease associated complex subunits. There was also a large set of ribosomes within the extremely high scoring disease genes, both nuclear and mitochondrial, which would most likely have a high probability to be

disease causing in a patient due to their fundamental cellular function. Functional annotation of the extremely high score disease genes using DAVID [311] identified many significantly enriched annotations within the set of genes such as transport ($p = 1.3x10^{-3}$), methylation ($p = 8.5x10^{-5}$) and apoptosis ($p = 1.1x10^{-3}$).

The multiclass neural network predicted 232 recessive disease genes out of the remaining mitochondrial genes (*Figure 5.12*), 82% of these were assigned a high recessive class score (≥0.5) with 49% assigned an extremely high score (≥0.75). This set of extremely high scoring genes was extremely similar to the set of genes assigned a high disease score by the binary neural network, including all of the un-associated mitochondrial complex subunits such as COX7C and COQ3. Function annotation identified many enriched annotations within this set of genes including ribosomal proteins ($p = 1.9x10^{-16}$), transport ($p = 1.4x10^{-4}$) and aminotransferase ($p = 6.3x10^{-3}$). The set of genes assigned a high 'other' disease class score (≥0.9) contained two mitochondrially encoded OXPHOS complex subunits, MT-ND1 and MT-ND4. These were assigned to the 'other' disease class as the known mitochondrially inherited disease genes, such as MT-COX1, were included in the 'other' disease training set. The highly predicted 'other' disease set of genes also contained a couple of ribosomes ($p = 4.4x10^{-3}$) and shared many similar significantly enriched annotations with the highly predicted recessive disease set of genes. The three mitochondrial genes predicted dominant with a high dominant class score (≥ 0.5), and low score for all other classes (≤0.4), were RPL18A and RPS8 which are both nuclear ribosomes. The nuclear ribosome RPS15A is a known autosomal dominant disease-causing gene [312] which is most likely the reason for their predictions. The low performance of the multiclass network on the dominant training set means that this prediction should not be taken in high confidence but does suggest that these genes do not look similar to any of the know recessive genes.

The distribution of the disease class score predictions for both networks on the t-SNE plot showed how similar the disease predictions were by the two networks (*Figure 5.13*). The set of genes predicted to be disease associated by both networks were also assigned a high disease class score (≥ 0.5) in both networks (*Figure 5.14*). The set of genes predicted to be disease and assigned an extremely high disease score (≥0.95) in both networks contained many genes which have already been

associated with diseases, including many of the mitochondrial complex subunits and the following discussed genes. ACTB has been associated with Baraitser-Winter syndrome which has symptoms such as growth problems and seizures [313]. YWHAZ has been associated with Popov-Chang syndrome, a neurodevelopmental disorder [314]. GNB1 which has been associated with mental retardation, seizures, hypotonia and neurodevelopmental problems [315]. CRYAB which has been associated with cardiomyopathy and myopathy [316]. All these symptoms have a large cross-over with common mitochondrial disease symptoms giving support for their prediction as potential mitochondrial disease genes. In addition, two well-known viral onco-genes were assigned an extremely high disease score by both networks, ATK1 [317] and SRC [318]. Many heat shock proteins, which are chaperone proteins and regulators of cellular adaptations to stress, such as HSPA8 and HSPA5 were assigned extremely high disease score by both networks. The heat shock protein HSPD1 is an already known mitochondrial disease gene [319] which gives support for their assignment and highlights these as an interesting family of proteins to investigate further. Many ribosomes were also predicted in this set including MRPS25 which is the only experimentally confirmed recessive mitochondrial disease gene to be identified since the training of both networks [320]. The gene was predicted to be disease in the binary network with an assigned disease class score of 0.78, and recessive with an assigned recessive class score of 0.72 by the multiclass network. Both networks therefore correctly predicted the disease association and disease type with relatively high class scores which, although only a single entry, gives experimentally verified support for both neural network predictions.

5.5. Conclusion

The features generated for the input data array identified properties of mitochondrial disease genes which could be used to linearly separate the known disease genes, the positive training set, and the negative training set. Combined into a single data array, the features were suitable for use in training a neural network for classification of mitochondrial disease with the aim of generating disease association scores for every gene in the predicted mitochondrial proteome for use in improving patient diagnosis rates. The binary neural network was successfully trained for binary classification of disease or non-disease with no sign of overfitting. The multiclass neural network showed moderate success but was ultimately limited by the small subset of highly variable known dominant diseases. However, the multiclass network showed great success in accurately assigning disease or non-disease status to the training set, giving it validity in predicting mitochondrial diseases vs non-disease. The two networks showed similarity in their predictions with many of the agreed upon highly predicted disease genes having already suspected disease associations. The networks also accurately predicted the only experimentally verified mitochondrial disease gene to be identified since the development and training of the two networks, the recessive disease gene MRPS25, a nuclear ribosome. The predicted disease association scores are currently being used in an NGS analysis pipeline at the MRC Mitochondrial Biology Unit for diagnosing mitochondrial disease patient sample

Chapter 6

Conclusions

6.1. Metabolic adaptations to mitochondrial dysfunction

The aim of this project was to investigate the metabolic adaptations to mitochondrial dysfunction in mammals. The large-scale metabolomics data set analysed in *Chapter 2* provided the means to begin identifying the metabolic adaptations to mitochondrial dysfunction caused by drug-induced mitochondrial complex III inhibition.

The combined analysis of hypothesis testing, support vector machines with recursive feature elimination and partial least squared discriminant analysis enabled the identification of metabolites found significantly different during mild mitochondrial dysfunction in mammals across all three methods. Hypothesis testing was also utilized to identify significantly different metabolites during high levels of mitochondrial dysfunction, the low biological replicate number prevented the use of the more complex analysis methods. The significant metabolites for both mild and high levels of mitochondrial dysfunction identified many perturbed pathways and metabolites which have important biological functions in relation to mitochondrial function and could be used as biomarkers, e.g. Serine. The analysis of the metabolomics data set revealed erroneous pathways which were difficult to explain based on current understanding of mitochondrial dysfunction. Signs of erroneous fatty acid β-oxidation were present in both mild and high levels of mitochondrial dysfunction, based on the significance of multiple 3-hydroxy fatty acids, as were signs of erroneous branched chain amino acid (BCAA) metabolism based on the presence of many hallmarks of known BCAA metabolism in-born diseases.

To investigate the potential causes for many of the identified significant metabolites and to provide a means to investigate mitochondrial dysfunction caused by other complex inhibitions, a multi-organ model of mitochondrial metabolism was developed as described in *Chapter 3*. The simulations of liver mitochondrial complex III inhibition predicted similar behaviour to the adaptations identified in the metabolomics data set, supporting its applicability in predicting *in vivo* outcomes. The simulations also provided a potential reasoning for the identified erroneous fatty acid β-oxidation and BCAA metabolism *in vivo* related to their by production of quinones, a highly undesirable behaviour in the model under complex III inhibition. The simulations of the other complex's inhibitions identified many pathways which could be exploited as unique biomarkers for each type of complex inhibition and provide a means in which to estimate the level of inhibition occurring. For example, complex I inhibition identified an inflection point in the behaviour of the conversion of pyruvate into lactate during only a small window of complex I inhibition levels which, if seen *in vivo*, would make an excellent biomarker and allow for the estimation of the complex I inhibition level occurring.

Future metabolomics studies should be performed on mitochondrial dysfunction caused by other types of complex inhibition. By comparing the adaptations occurring *in vivo* to different types of mitochondrial dysfunction with the adaptations identified in *Chapter 2* would enable the identification of unique high confidence biomarkers of complex III inhibition, along with high confidence biomarkers of overall mitochondrial dysfunction. In addition, the multi-organ complex inhibition simulations provide a means in which to direct future experimentation. Direct testing of the predictions of the simulations by targeted assays *in vivo* would provide higher confidence in the model and identify potential biomarkers of mitochondrial dysfunction. Given successful predictions, the model could then be used as part of future computational studies investigating other types of mitochondrial dysfunction, such as the inhibition of specific transport proteins.

6.2. Predicting novel mitochondrial disease genes

The aim of this project was to generate a meaningful way in which a suspected mitochondrial disease patient's variant could be filtered through to prioritize candidate disease-causing genes for experimental verification. The necessary requirement to filtering variants for mitochondrial disease-causing genes is the mitochondrial proteome, which is currently considered incomplete. Therefore, the first step to achieving the aim of this project was to expand on the current mitochondrial proteome using the vast amounts of published data on mitochondrial protein localisation.

The MitoMiner database contained a wealth of manually curated computational and experimental protein localisation data sources. The data sources were used to train a support vector machine (SVM) to predict protein mitochondrial localisation as described in *Chapter 4*. The tuned SVM showed no signs of over or underfitting and reported a high generalisation accuracy of 92.2%. A total of 442 novel proteins were predicted highly likely to be mitochondrially localised which expanded the current mitochondrial proteome to 1,626 proteins. These novel proteins can now be used to investigate mitochondrial diseases and the predictions on the remaining human proteome can be used as a means to evaluate current evidence on each protein's mitochondrial localisation. Future work can continue to improve the predictions by including up-to-date data sources in the input array used for training the SVM and by updating the protein training sets as new proteins are experimentally verified.

Having established a predicted mitochondrial proteome, the proteome could then be used as the basis in which to investigate predicting novel mitochondrial disease associated genes. Two neural networks were trained to predict disease association and disease type of each gene in the predicted mitochondrial proteome as described in *Chapter 5*. The neural networks were trained using features generated from multiple different types of data including gene tissue expression, protein-protein interaction network metrics, population study metrics and gene evolutionary history.

Both networks showed great success in accurately predicting the disease or non-disease status of the positive and negative training sets. One of the networks was trained for binary classification of disease or non-disease which achieved high

performance metrics, no sign of overfitting and a generalisation accuracy of 90.18%. The second network was trained for multiclass classification of non-disease, recessive disease, dominant disease or 'other' disease. The network achieved reasonable performance metrics but was severely hindered by the dominant diseases as there is currently only a small set of known dominant disease genes which had a high in-group variance and high similarity to the recessive disease genes. The multiclass network had a reasonable generalisation accuracy of 81.25% and performed similarly well to the binary neural network when only considering its predictions as disease or non-disease i.e. combining the dominant, recessive and 'other' disease predictions into a single disease prediction.

The two neural networks agreed upon 304 predicted mitochondrial disease-associated genes within the predicted mitochondrial proteome. This set of proteins, and the complete list of predicted disease-association probabilities by both neural networks, can be used to filter through a patient's list of variants for candidate disease-causing genes. The predictions by the two neural networks successfully predicted the only experimentally verified new mitochondrial disease gene published since the training of the networks was completed.

Future work to improve the networks could be performed by organising the 'other' disease group into either dominant or recessive which would simplify the multiclass classification problem by reducing the number of classes and increase the dominant training set size. Generating a feature based on a gene's chromosome may provide a way in which to address the issue of having X-linked diseases separated into dominant and recessive. This would also allow for a dummy variable to be generated in the same feature for mtDNA diseases, although they would ultimately have to be classified as dominant or recessive. The framework for training the neural networks has been performed in this project which enables future work to improve upon the networks by adding more features to the input training array and updating the training sets as new genes are identified.

6.3. Thesis summary

In summary, in this thesis I have begun addressing two major issues related to mitochondria in toxicology and mitochondrial diseases. The metabolic adaptations identified in the metabolomics study provide the first steps in trying to understand mitochondrial dysfunction at a metabolic level which can be built upon greatly by future experimental studies. The predictions by the modelling of complex inhibition provide a basis for directing future experiments and if proven correct, could provide *in-silico* predictions for multiple other types of mitochondrial dysfunction relevant for both toxicology and mitochondrial disease. The predicted mitochondrial proteome collates multiple protein localisation data sources into a single, easy to understand value and identified many novel mitochondrial proteins which can be used as a basis for investigating mitochondrial disease. Predicted disease-association probabilities of the predicted mitochondrial proteome provide a means for filtering a patient's variants for candidate disease-causing genes which was a previously difficult and time-consuming process. The framework for generating the predictions of protein localisation and disease-association has been generated which can be iteratively improved upon in future as new information gets published. The predicted mitochondrial proteome and disease-association probabilities are currently being used in an NGS pipeline for diagnosing mitochondrial disease patient samples in the MRC Mitochondrial Biology Unit.

References

[1] Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Molecular Biology of the Cell 2002.

[2] Embley MT, Martin W. Eukaryotic evolution, changes and challenges. Nature 2006;440:623–30. doi:10.1038/nature04546 .

[3] Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Pontén T, Alsmark CU, Podowski RM, et al. The genome sequence of Rickettsia prowazekii and the origin of mitochondria. Nature 1998;396:24094. doi:10.1038/24094 .

[4] Gribaldo S, Poole AM, Daubin V, Forterre P, Brochier-Armanet C. The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse? Nat Rev Microbiol 2010;8:743. doi:10.1038/nrmicro2426 .

[5] Sagan L. On the origin of mitosing cells. J Theor Biol 1967;14:225-IN6. doi:10.1016/0022-5193(67)90079-3 .

[6] Youle RJ, van der Bliek AM. Mitochondrial Fission, Fusion, and Stress. Science 2012;337:1062–5. doi:10.1126/science.1219855 .

[7] Longo DL, Archer SL. Mitochondrial Dynamics — Mitochondrial Fission and Fusion in Human Diseases. New Engl J Medicine 2013;369:2236–51. doi:10.1056/nejmra1215233 .

[8] Dolman NJ, Gerasimenko JV, Gerasimenko OV, Voronina SG, Petersen OH, Tepikin AV. Stable Golgi-Mitochondria Complexes and Formation of Golgi Ca2+ Gradients in Pancreatic Acinar Cells. J Biol Chem 2005;280:15794–9. doi:10.1074/jbc.m412694200 .

[9] Giacomello M, Pellegrini L. The coming of age of the mitochondria–ER contact: a matter of thickness. Cell Death Differ 2016;23:1417–27. doi:10.1038/cdd.2016.52 .

[10] Colombini M. VDAC: The channel at the interface between mitochondria and the cytosol. Mol Cell Biochem 2004;256–257:107–15. doi:10.1023/b:mcbi.0000009862.17396.8d .

[11] Kunji E. The role and structure of mitochondrial carriers. Febs Lett 2004;564:239–44. doi:10.1016/s0014-5793(04)00242-x .

[12] Zick M, Rabl R, Reichert AS. Cristae formation—linking ultrastructure and function of mitochondria. Biochimica Et Biophysica Acta Bba - Mol Cell Res 2009;1793:5–19. doi:10.1016/j.bbamcr.2008.06.013 .

[13] Cogliati S, Frezza C, Soriano M, Varanita T, Quintana-Cabrera R, Corrado M, et al. Mitochondrial Cristae Shape Determines Respiratory Chain Supercomplexes Assembly and Respiratory Efficiency. Cell 2013;155:160–71. doi:10.1016/j.cell.2013.08.032 .

[14] Yoshida M, Muneyuki E, Hisabori T. ATP synthase — a marvellous rotary engine of the cell. Nat Rev Mol Cell Bio 2001;2:35089509. doi:10.1038/35089509 .

[15] Watt IN, Montgomery MG, Runswick MJ, Leslie AG, Walker JE. Bioenergetic cost of making an adenosine triphosphate molecule in animal mitochondria. Proc National Acad Sci 2010;107:16823–7. doi:10.1073/pnas.1011099107 .

[16] Zhu J, Vinothkumar KR, Hirst J. Structure of mammalian respiratory complex I. Nature 2016;536:354. doi:10.1038/nature19095 .

[17] Crofts AR, Berry EA. Structure and function of the cytochrome bc1 complex of mitochondria and photosynthetic bacteria. Curr Opin Struc Biol 1998;8:501–9. doi:10.1016/s0959-440x(98)80129-2 .

[18] Capaldi RA. Structure and assembly of cytochrome c oxidase. Arch Biochem Biophys 1990;280:252–62. doi:10.1016/0003-9861(90)90327-u .

[19] Sun F, Huo X, Zhai Y, Wang A, Xu J, Su D, et al. Crystal Structure of Mitochondrial Respiratory Membrane Protein Complex II. Cell 2005;121:1043–57. doi:10.1016/j.cell.2005.05.025 .

[20] Akram M. Citric Acid Cycle and Role of its Intermediates in Metabolism. Cell Biochem Biophys 2014;68:475–8. doi:10.1007/s12013-013-9750-1 .

[21] Miller WL. Steroid hormone synthesis in mitochondria. Mol Cell Endocrinol 2013;379:62–73. doi:10.1016/j.mce.2013.04.014 .

[22] Ryter SW, Tyrrell RM. The heme synthesis and degradation pathways: role in oxidant sensitivity Heme oxygenase has both pro- and antioxidant properties. Free Radical Bio Med 2000;28:289–309. doi:10.1016/s0891-5849(99)00223-3 .

[23] Stehling O, Lill R. The Role of Mitochondria in Cellular Iron–Sulfur Protein Biogenesis: Mechanisms, Connected Processes, and Diseases. Csh Perspect Biol 2013;5:a011312. doi:10.1101/cshperspect.a011312 .

[24] Contreras L, Drago I, Zampese E, Pozzan T. Mitochondria: The calcium connection. Biochimica Et Biophysica Acta Bba - Bioenergetics 2010;1797:607–18. doi:10.1016/j.bbabio.2010.05.005 .

[25] Wang C, Youle RJ. The Role of Mitochondria in Apoptosis*. Genetics 2009;43:95–118. doi:10.1146/annurev-genet-102108-134850 .

[26] Zou H, Li Y, Liu X, Wang X. An APAF-1·Cytochrome c Multimeric Complex Is a Functional Apoptosome That Activates Procaspase-9. J Biol Chem 1999;274:11549–56. doi:10.1074/jbc.274.17.11549 .

[27] Anderson S, Bankier A, Barrell B, de Bruijn M, Coulson A, Drouin J, et al. Sequence and organization of the human mitochondrial genome. Nature 1981;290:457–65. doi:10.1038/290457a0 .

[28] Meisinger C, Sickmann A, Pfanner N. The Mitochondrial Proteome: From Inventory to Function. Cell 2008;134:22–4. doi:10.1016/j.cell.2008.06.043 .

[29] Smith AC, Robinson AJ. MitoMiner v4.0: an updated database of mitochondrial localization evidence, phenotypes and diseases. Nucleic Acids Res 2018;47:gky1072-. doi:10.1093/nar/gky1072 .

[30] Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science 2015;347:1260419. doi:10.1126/science.1260419 .

[31] Itzhak DN, Tyanova S, Cox J, Borner GH. Global, quantitative and dynamic mapping of protein subcellular localization. Elife 2016;5:e16950. doi:10.7554/elife.16950 .

[32] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry MJ, et al. Gene Ontology: tool for the unification of biology. Nat Genet 2000;25:25–9. doi:10.1038/75556 .

[33] Fukasawa Y, Tsuji J, Fu S-C, Tomii K, Horton P, Imai K. MitoFates: Improved Prediction of Mitochondrial Targeting Sequences and Their Cleavage Sites. Mol Cell Proteomics 2015;14:1113–26. doi:10.1074/mcp.m114.043083 .

[34] Fernando V, Guda P, Subramaniam S, Guda C. Cardiovascular Proteomics, Methods and Protocols 2006:375–83. doi:10.1385/1-59745-214-9:375 .

[35] Elstner M, Andreoli C, Klopstock T, Meitinger T, Prokisch H. Chapter 1 The Mitochondrial Proteome Database MitoP2. Methods Enzymol 2009;457:3–20. doi:10.1016/s0076-6879(09)05001-0 .

[36] Pagliarini DJ, Calvo SE, Chang B, Sheth SA, Vafai SB, Ong S-E, et al. A Mitochondrial Protein Compendium Elucidates Complex I Disease Biology. Cell 2008;134:112–23. doi:10.1016/j.cell.2008.06.016 .

[37] Calvo SE, Clauser KR, Mootha VK. MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. Nucleic Acids Res 2016;44:D1251–7. doi:10.1093/nar/gkv1003 .

[38] Winklhofer KF, Haass C. Mitochondrial dysfunction in Parkinson's disease. Biochimica Et Biophysica Acta Bba - Mol Basis Dis 2010;1802:29–44. doi:10.1016/j.bbadis.2009.08.013 .

[39] Varastet M, Riche D, Maziere M, Hantraye P. Chronic MPTP treatment reproduces in baboons the differential vulnerability of mesencephalic dopaminergic neurons observed in parkinson's disease. Neuroscience 1994;63:47–56. doi:10.1016/0306-4522(94)90006-x .

[40] Moratalla R, Quinn B, DeLanney L, Irwin I, Langston J, Graybiel A. Differential vulnerability of primate caudate-putamen and striosome-matrix dopamine systems to the neurotoxic effects of 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine. Proc National Acad Sci 1992;89:3859–63. doi:10.1073/pnas.89.9.3859 .

[41] Pickrell AM, Youle RJ. The Roles of PINK1, Parkin, and Mitochondrial Fidelity in Parkinson's Disease. Neuron 2015;85:257–73. doi:10.1016/j.neuron.2014.12.007 .

[42] Vives-Bauza C, Zhou C, Huang Y, Cui M, de Vries R, Kim J, et al. PINK1-dependent recruitment of Parkin to mitochondria in mitophagy. Proc National Acad Sci 2010;107:378–83. doi:10.1073/pnas.0911187107 .

[43] Johri A, Beal FM. Mitochondrial Dysfunction in Neurodegenerative Diseases. J Pharmacol Exp Ther 2012;342:619–30. doi:10.1124/jpet.112.192138 .

[44] Chuang DT. Maple syrup urine disease: It has come a long way. J Pediatrics 1998;132:S17–23. doi:10.1016/s0022-3476(98)70523-2 .

[45] Nellis MM, Danner DJ. Gene Preference in Maple Syrup Urine Disease. Am J Hum Genetics 2001;68:232–7. doi:10.1086/316950 .

[46] Rossignol R, Faustin B, Rocher C, Malgat M, Mazat J-P, Letellier T. Mitochondrial threshold effects. Biochem J 2003;370:751–62. doi:10.1042/bj20021594 .

[47] Neustadt J, Pieczenik SR. Medication-induced mitochondrial damage and disease. Mol Nutr Food Res 2008;52:780–8. doi:10.1002/mnfr.200700075 .

[48] Apostolova N, Blas-García A, Esplugues JV. Mitochondrial interference by anti-HIV drugs: mechanisms beyond Pol-γ inhibition. Trends Pharmacol Sci 2011;32:715–25. doi:10.1016/j.tips.2011.07.007 .

[49] Finsterer J, Segall L. Drugs interfering with mitochondrial disorders. Drug Chem Toxicol 2009;33:138–51. doi:10.3109/01480540903207076 .

[50] Szewczyk A, Wojtczak L. Mitochondria as a Pharmacological Target. Pharmacol Rev 2002;54:101–27. doi:10.1124/pr.54.1.101 .

[51] Barbosa IA, Machado NG, Skildum AJ, Scott PM, Oliveira PJ. Mitochondrial remodeling in cancer metabolism and survival: Potential for new therapies. Biochimica Et Biophysica Acta Bba - Rev Cancer 2012;1826:238–54. doi:10.1016/j.bbcan.2012.04.005 .

[52] Finsterer J. Treatment of mitochondrial disorders. Eur J Paediatr Neuro 2010;14:29–44. doi:10.1016/j.ejpn.2009.07.005 .

[53] Jafarian I, Eskandari M, Mashayekhi V, Ahadpour M, Hosseini M-J. Toxicity of valproic acid in isolated rat liver mitochondria. Toxicol Mech Method 2013;23:617–23. doi:10.3109/15376516.2013.821567 .

[54] Liberti MV, Locasale JW. The Warburg Effect: How Does it Benefit Cancer Cells? Trends Biochem Sci 2016;41:211–8. doi:10.1016/j.tibs.2015.12.001 .

[55] Fulda S, Galluzzi L, Kroemer G. Targeting mitochondria for cancer therapy. Nat Rev Drug Discov 2010;9:447. doi:10.1038/nrd3137 .

[56] Nihei C, Fukai Y, Kita K. Trypanosome alternative oxidase as a target of chemotherapy. Biochimica Et Biophysica Acta Bba - Mol Basis Dis 2002;1587:234–9. doi:10.1016/s0925-4439(02)00086-8 .

[57] Freneaux E, Fromenty B, Berson A, Labbe G, Degott C, Letteron P, et al. Stereoselective and nonstereoselective effects of ibuprofen enantiomers on mitochondrial beta-oxidation of fatty acids. J Pharmacol Exp Ther 1990;255:529–35.

[58] Morgan S, Grootendorst P, Lexchin J, Cunningham C, Greyson D. The cost of drug development: A systematic review. Health Policy 2011;100:4–17. doi:10.1016/j.healthpol.2010.12.002 .

[59] Waring MJ, Arrowsmith J, Leach AR, Leeson PD, Mandrell S, Owen RM, et al. An analysis of the attrition of drug candidates from four major pharmaceutical companies. Nat Rev Drug Discov 2015;14:nrd4609. doi:10.1038/nrd4609 .

[60] Kramer JA, Sagartz JE, Morris DL. The application of discovery toxicology and pathology towards the design of safer pharmaceutical lead candidates. Nat Rev Drug Discov 2007;6:636–49. doi:10.1038/nrd2378 .

[61] Innovation or Stagnation: Challenge and opportunity on the critical path to new medical products. Washington DC, USA: Food and Drug Administration. 2004.

[62] Marroquin LD, Hynes J, Dykens JA, Jamieson JD, Will Y. Circumventing the Crabtree Effect: Replacing Media Glucose with Galactose Increases Susceptibility of HepG2 Cells to Mitochondrial Toxicants. Toxicol Sci 2007;97:539–47. doi:10.1093/toxsci/kfm052 .

[63] Dykens JA, Will Y. The significance of mitochondrial toxicity testing in drug development. Drug Discov Today 2007;12:777–85. doi:10.1016/j.drudis.2007.07.013 .

[64] Sansbury BE, Jones SP, Riggs DW, Darley-Usmar VM, Hill BG. Bioenergetic

function in cardiovascular cells: The importance of the reserve capacity and its biological regulation. Chem-Biol Interact 2011;191:288–95. doi:10.1016/j.cbi.2010.12.002 .

[65] Desler C, Hansen T, Frederiksen J, Marcker M, Singh KK, Rasmussen L. Is There a Link between Mitochondrial Reserve Respiratory Capacity and Aging? J Aging Res 2012;2012:192503. doi:10.1155/2012/192503 .

[66] Strimbu K, Tavel JA. What are biomarkers? Curr Opin Hiv Aids 2010;5:463–6. doi:10.1097/coh.0b013e32833ed177 .

[67] Barry MJ. Prostate-Specific–Antigen Testing for Early Diagnosis of Prostate Cancer. New Engl J Medicine 2001;344:1373–7. doi:10.1056/nejm200105033441806 .

[68] Horgan RP, Kenny LC. ‘Omic’ technologies: genomics, transcriptomics, proteomics and metabolomics. Obstetrician Gynaecol 2011;13:189–95. doi:10.1576/toag.13.3.189.27672 .

[69] Dettmer K, Aronov PA, Hammock BD. Mass spectrometry-based metabolomics. Mass Spectrom Rev 2007;26:51–78. doi:10.1002/mas.20108 .

[70] Wishart DS. Quantitative metabolomics using NMR. Trac Trends Anal Chem 2008;27:228–37. doi:10.1016/j.trac.2007.12.001 .

[71] Monteiro, Carvalho M, Bastos M, de Pinho GP. Metabolomics Analysis for Biomarker Discovery: Advances and Challenges. Curr Med Chem 2013;20:257–71. doi:10.2174/092986713804806621 .

[72] Spratlin JL, Serkova NJ, Eckhardt GS. Clinical Applications of Metabolomics in Oncology: A Review. Clin Cancer Res 2009;15:431–40. doi:10.1158/1078-0432.ccr-08-1059 .

[73] Wishart DS, Jewison T, Guo A, Wilson M, Knox C, Liu Y, et al. HMDB 3.0—The

Human Metabolome Database in 2013. Nucleic Acids Res 2013;41:D801–7. doi:10.1093/nar/gks1065 .

[74] Patti GJ, Yanes O, Siuzdak G. Innovation: Metabolomics: the apogee of the omics trilogy. Nat Rev Mol Cell Bio 2012;13:263. doi:10.1038/nrm3314 .

[75] Legault J, Strittmatter L, Tardif J, Sharma R, Tremblay-Vaillancourt V, Aubut C, et al. A Metabolic Signature of Mitochondrial Dysfunction Revealed through a Monogenic Form of Leigh Syndrome. Cell Reports 2015;13:981–9. doi:10.1016/j.celrep.2015.09.054 .

[76] van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. Bmc Genomics 2006;7:142. doi:10.1186/1471-2164-7-142 .

[77] Worley B, Powers R. Multivariate Analysis in Metabolomics. Curr Metabolomics 2012;1:92–107. doi:10.2174/2213235x11301010092 .

[78] Bogdanov M, Matson WR, Wang L, Matson T, Saunders-Pullman R, Bressman SS, et al. Metabolomic profiling to develop blood biomarkers for Parkinson's disease. Brain 2008;131:389-96. doi: 10.1093/brain/awm304 .

[79] Zhang J, Bowers J, Liu L, Wei S, Gowda GA, Hammoud Z, et al. Esophaheal cancer metabolite biomerks detected by LC-MS and NMR methods. PLoS One 2012;7: e30181. doi: 10.1371/journal.pone.0030181 .

[80] Brereton RG, Lloyd GR. Partial least squares discriminant analysis: taking the magic away. J Chemometr 2014;28:213–25. doi:10.1002/cem.2609 .

[81] Rubingh CM, Bijlsma S, Derks EP, Bobeldijk I, Verheij ER, Kochhar S, et al. Assessing the performance of statistical validation tools for megavariate metabolomics data. Metabolomics 2006;2:53–61. doi:10.1007/s11306-006-0022-6 .

[82] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–97. doi:10.1007/bf00994018 .

[83] Mahadevan S, Shah SL, Marrie TJ, Slupsky CM. Analysis of Metabolomic Data Using Support Vector Machines. Anal Chem 2008;80:7562–70. doi:10.1021/ac800954c .

[84] Heinemann J, Mazurie A, Tokmina-Lukaszewska M, Beilman GJ, Bothner B. Application of support vector machines to metabolomics experiments with limited replicates. Metabolomics 2014;10:1121–8. doi:10.1007/s11306-014-0651-0 .

[85] Guan W, Zhou M, Hampton CY, Benigno BB, Walker LD, Gray A, et al. Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines. BMC Bioinformatics 2009;10:259. doi: 10.1186/1471-2105-10-259 .

[86] Bertini I, Calabro A, De Carli V, Luchinat C, Nepi S, Porfirio B, et al. The metabonomic signature of celiac disease. J Proteome Res 2009;8:170-7. doi: 10.1021/pr800548z .

[87] Broom AJ, Ambroso J, Brunori G, Burns AK, Armitage JR, Francis I, et al. Effects of mid-respiratory chain inhibition on mitochondrial function in vitro and in vivo. Toxicol Res-Uk 2015;5:136–50. doi:10.1039/c5tx00197h .

[88] MATLAB version 8.6.0, R2015b. Natick, Massachusetts: The MathWorks Inc., 2015. https://www.mathworks.com/ .

[89] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2018. https://www.R-project.org/ .

[90] Welch B. The generalization of ‘student’s’ problem when several different population varlances are involved. Biometrika 1947;34:28–35. doi:10.1093/biomet/34.1-2.28 .

[91] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J Royal Statistical Soc Ser B Methodol 1995;57:289–300. doi:10.1111/j.2517-6161.1995.tb02031.x .

[92] Python Core Team. Python: A dynamic, open source programming language. Python Software Foundation. 2019. https://www.python.org/ .

[93] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0--Fundamental Algorithms for Scientific Computing in Python 2019.

[94] Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python n.d.

[95] Spearman C. The Proof and Measurement of Association between Two Things. Am J Psychology 1904;15:72. doi:10.2307/1412159 .

[96] Guyon I, Weston J, Barnhill S, Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. Mach Learn 2002;46:389–422. doi:10.1023/a:1012487302797 .

[97] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python 2012.

[98] PLS Toolbox. Eigenvector Research, Inc. 2015. https://www.eigenvector.com/software/pls_toolbox.htm .

[99] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res 2003;13:2498–504. doi:10.1101/gr.1239303 .

[100] Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 2000;28:27–30. doi:10.1093/nar/28.1.27 .

[101] Henseler J, Ringle CM, Sinkovics RR. New Challenges to International

Marketing. Adv Int Marketing 2009;Volume 20:277–319. doi:10.1108/s1474-7979(2009)0000020014 .

[102] Bergoffen J, Kaplan P, Hale D, Bennett M, Berry G. Marked elevation of urinary 3-hydroxydecanedioic acid in a malnourished infant with glycogen storage disease, mimicking long-chainl-3-hydroxyacyl-CoA dehydrogenase deficiency. J Inherit Metab Dis 1993;16:851–6. doi:10.1007/bf00714277 .

[103] Bennett M, Weinberger M, Sherwood W, Burlina A. Secondary 3-hydroxydicarboxylic aciduria mimicking long-chain 3-hydroxyacyl-CoA dehydrogenase deficiency. J Inherit Metab Dis 1994;17:283–6. doi:10.1007/bf00711808 .

[104] Millington DS, Stevens RD. Metabolic Profiling, Methods and Protocols. Methods Mol Biology Clifton N J 2010;708:55–72. doi:10.1007/978-1-61737-985-7_3.

[105] Xu G, Hansen J, Zhao X, Chen S, Hoene M, Wang X, et al. Liver and Muscle Contribute Differently to the Plasma Acylcarnitine Pool During Fasting and Exercise in Humans. J Clin Endocrinol Metabolism 2016;101:5044–52. doi:10.1210/jc.2016-1859 .

[106] Stanley CA, Hale DE, Berry GT, Deleeuw S, Boxer J, Bonnefont J-P. A Deficiency of Carnitine–Acylcarnitine Translocase in the Inner Mitochondrial Membrane. New Engl J Medicine 1992;327:19–23. doi:10.1056/nejm199207023270104 .

[107] Haas RH, Parikh S, Falk MJ, Saneto RP, Wolf NI, et al. The in-depth evaluation of suspected mitochondrial disease. Mol Genet Metab 2008;94:16–37. doi:10.1016/j.ymgme.2007.11.018 .

[108] Taylor L, Curthoys NP. Glutamine metabolism: Role in acid-base balance*. Biochem Mol Biol Edu 2004;32:291–304. doi:10.1002/bmb.2004.494032050388 .

[109] Smith RJ. Glutamine Metabolism and Its Physiologic Importance. Jpen-Parenter Enter 1990;14:40S-44S. doi:10.1177/014860719001400402 .

[110] Yang C, Ko B, Hensley CT, Jiang L, Wasti AT, Kim J, et al. Glutamine Oxidation Maintains the TCA Cycle and Cell Survival during Impaired Mitochondrial Pyruvate Transport. Mol Cell 2014;56:414–24. doi:10.1016/j.molcel.2014.09.025 .

[111] Kalhan SC, Hanson RW. Resurgence of Serine: An Often Neglected but Indispensable Amino Acid. J Biol Chem 2012;287:19786–91. doi:10.1074/jbc.r112.357194 .

[112] Bao X, Ong S-E, Goldberger O, Peng J, Sharma R, Thompson DA, et al. Mitochondrial dysfunction remodels one-carbon metabolism in human cells. Elife 2016;5:e10575. doi:10.7554/elife.10575 .

[113] Yoshida T, Kikuchi G. Comparative Study on Major Pathways of Glycine and Serine Catabolism in Vertebrate Livers*. J Biochem 1972;72:1503–16. doi:10.1093/oxfordjournals.jbchem.a130042 .

[114] House JD, Hall BN, Brosnan JT. Threonine metabolism in isolated rat hepatocytes. Am J Physiol-Endoc M 2001;281:E1300–7. doi:10.1152/ajpendo.2001.281.6.e1300 .

[115] Stein L, Imai S. The dynamic regulation of NAD metabolism in mitochondria. Trends Endocrinol Metabolism 2012;23:420–8. doi:10.1016/j.tem.2012.06.005 .

[116] Fan J, Ye J, Kamphorst JJ, Shlomi T, Thompson CB, Rabinowitz JD. Quantitative flux analysis reveals folate-dependent NADPH production. Nature 2014;510:298. doi:10.1038/nature13236 .

[117] Saenger W. Structure and Function of Nucleosides and Nucleotides. Angewandte Chemie Int Ed Engl 1973;12:591–601. doi:10.1002/anie.197305911 .

[118] Shyh-Chang N, Locasale JW, Lyssiotis CA, Zheng Y, Teo R, Ratanasirintrawoot S, et al. Influence of Threonine Metabolism on *S*-Adenosylmethionine and Histone Methylation. Science 2013;339:222–6. doi:10.1126/science.1226603 .

[119] Ying W. NAD/NADH and NADP/NADPH in Cellular Functions and Cell Death Regulation and Biological Consequences. Antioxid Redox Sign 2008;10:179–206. doi:10.1089/ars.2007.1672 .

[120] Forman H, Zhang H, Rinna A. Glutathione: Overview of its protective roles, measurement, and biosynthesis. Mol Aspects Med 2009;30:1–12. doi:10.1016/j.mam.2008.08.006 .

[121] Murphy MP. How mitochondria produce reactive oxygen species. Biochem J 2009;417:1–13. doi:10.1042/bj20081386 .

[122] Murphy MP. Mitochondrial Dysfunction Indirectly Elevates ROS Production by the Endoplasmic Reticulum. Cell Metab 2013;18:145–6. doi:10.1016/j.cmet.2013.07.006 .

[123] Owen JB, Butterfield AD. Protein Misfolding and Cellular Stress in Disease and Aging, Concepts and Protocols. Methods Mol Biology Clifton N J 2010;648:269–77. doi:10.1007/978-1-60761-756-3_18 .

[124] Bachhawat A, Yadav S. The glutathione cycle: Glutathione metabolism beyond the γ-glutamyl cycle. Iubmb Life 2018;70:585–92. doi:10.1002/iub.1756 .

[125] Gao X, Lee K, Reid MA, Sanderson SM, Qiu C, Li S, et al. Serine Availability Influences Mitochondrial Dynamics and Function through Lipid Metabolism. Cell Reports 2018;22:3507–20. doi:10.1016/j.celrep.2018.03.017 .

[126] Colbeau A, Nachbaur J, Vignais PM. Enzymac characterization and lipid composition of rat liver subcellular membranes. Biochimica Et Biophysica Acta Bba - Biomembr 1971;249:462–92. doi:10.1016/0005-2736(71)90123-4 .

[127] Pelech S, Pritchard P, Brindley D, Vance D. Fatty acids promote translocation of CTP:phosphocholine cytidylyltransferase to the endoplasmic reticulum and stimulate rat hepatic phosphatidylcholine synthesis. J Biological Chem 1983;258:6782–8.

[128] Cornell R, Vance DE. Translocation of CTP:phosphocholine cytidylyltransferase from cytosol to membranes in HeLa cells: stimulation by fatty acid, fatty alcohol, mono- and diacylglycerol. Biochimica Et Biophysica Acta Bba - Lipids Lipid Metabolism 1987;919:26–36. doi:10.1016/0005-2760(87)90214-1 .

[129] Osman C, Voelker DR, Langer T. Making heads or tails of phospholipids in mitochondria. J Cell Biology 2011;192:7–16. doi:10.1083/jcb.201006159 .

[130] Klein J. Membrane breakdown in acute and chronic neurodegeneration: focus on choline-containing phospholipids. J Neural Transm 2000;107:1027–63. doi:10.1007/s007020070051 .

[131] Gibellini F, Smith TK. The Kennedy pathway—De novo synthesis of phosphatidylethanolamine and phosphatidylcholine. Iubmb Life 2010;62:414–28. doi:10.1002/iub.337 .

[132] Fernández-Murray PJ, McMaster CR. Glycerophosphocholine Catabolism as a New Route for Choline Formation for Phosphatidylcholine Synthesis by the Kennedy Pathway. J Biol Chem 2005;280:38290–6. doi:10.1074/jbc.m507700200 .

[133] Cheng M, Bhujwalla ZM, Glunde K. Targeting Phospholipid Metabolism in Cancer. Frontiers Oncol 2016;6:266. doi:10.3389/fonc.2016.00266 .

[134] Ridgway ND. The role of phosphatidylcholine and choline metabolites to cell proliferation and survival. Crit Rev Biochem Mol 2013;48:20–38. doi:10.3109/10409238.2012.735643 .

[135] Kobayashi R, Shimomura Y, Murakami T, Nakai N, Otsuka M, Arakawa N, et

al. Hepatic Branched-Chain α-Keto Acid Dehydrogenase Complex in Female Rats: Activation by Exercise and Starvation. J Nutr Sci Vitaminol 1999;45:303–9. doi:10.3177/jnsv.45.303 .

[136] Kobayashi R, Murakami T, Obayashi M, Nakai N, Jaskiewicz J, Fujiwara Y, et al. Clofibric acid stimulates branched-chain amino acid catabolism by three mechanisms. Arch Biochem Biophys 2002;407:231–40. doi:10.1016/s0003-9861(02)00472-1 .

[137] Li T, Zhang Z, Kolwicz SC, Abell L, Roe ND, Kim M, et al. Defective Branched-Chain Amino Acid Catabolism Disrupts Glucose Metabolism and Sensitizes the Heart to Ischemia-Reperfusion Injury. Cell Metab 2017;25:374–85. doi:10.1016/j.cmet.2016.11.005 .

[138] Korman SH, Andresen BS, Zeharia A, Gutman A, Boneh A, Pitt JJ. 2-Ethylhydracrylic Aciduria in Short/Branched-Chain Acyl-CoA Dehydrogenase Deficiency: Application to Diagnosis and Implications for the R-Pathway of Isoleucine Oxidation. Clin Chem 2005;51:610–7. doi:10.1373/clinchem.2004.043265 .

[139] Bischof F, Nägele T, Wanders RJ, Trefz FK, Melms A. 3-hydroxy-3-methylglutaryl-CoA lyase deficiency in an adult with leukoencephalopathy. Ann Neurol 2004;56:727–30. doi:10.1002/ana.20280 .

[140] Wendel U, Becker K, Przyrembel H, Bulla M, Manegold C, Mench-Hoinowski A, et al. Peritoneal dialysis in maple-syrup-urine disease: Studies on branched-chain amino and keto acids. Eur J Pediatr 1980;134:57–63. doi:10.1007/bf00442404 .

[141] Viegas C, da Ferreira G, Schuck P, Tonin A, Zanatta Â, de Wyse A, et al. Evidence that 3-hydroxyisobutyric acid inhibits key enzymes of energy metabolism in cerebral cortex of young rats. Int J Dev Neurosci 2008;26:293–9. doi:10.1016/j.ijdevneu.2008.01.007 .

[142] Arinze IJ. Facilitating understanding of the purine nucleotide cycle and the one-

carbon pool: Part I: The purine nucleotide cycle. Biochem Mol Biol Edu 2005;33:165–8. doi:10.1002/bmb.2005.494033032469 .

[143] Zikánová M, Krijt J, Hartmannová H, Kmoch S. Preparation of 5-amino-4-imidazole-N-succinocarboxamide ribotide, 5-amino-4-imidazole-N-succinocarboxamide riboside and succinyladenosine, compounds usable in diagnosis and research of adenylosuccinate lyase deficiency. J Inherit Metab Dis 2005;28:493–9. doi:10.1007/s10545-005-0493-z .

[144] Phang JM, Liu W, Hancock CN, Fischer JW. Proline metabolism and cancer. Curr Opin Clin Nutr 2015;18:71–7. doi:10.1097/mco.0000000000000121 .

[145] Haack TB, Haberberger B, Frisch E-M, Wieland T, Iuso A, Gorza M, et al. Molecular diagnosis in mitochondrial complex I deficiency using exome sequencing. J Med Genet 2012;49:277. doi:10.1136/jmedgenet-2012-100846 .

[146] Jackson C, Nuoffer J-M, Hahn D, Prokisch H, Haberberger B, Gautschi M, et al. Mutations in SDHD lead to autosomal recessive encephalomyopathy and isolated mitochondrial complex II deficiency. J Med Genet 2014;51:170. doi:10.1136/jmedgenet-2013-101932 .

[147] Gaignard P, Eyer D, Lebigot E, Oliveira C, Therond P, Boutron A, et al. UQCRC2 mutation in a patient with mitochondrial complex III deficiency causing recurrent liver failure, lactic acidosis and hypoglycemia. J Hum Genet 2017;62:729. doi:10.1038/jhg.2017.22 .

[148] Biervliet VJ, Bruinvis L, Ketting D, Bree P, Heiden VC, Wadman S, et al. Hereditary mitochondrial myopathy with lactic acidemia, a De Toni-Fanconi-Debré syndrome, and a defective respiratory chain in voluntary striated muscles. Pediatr Res 1977;11:1088–93. doi:10.1203/00006450-197711100-00005 .

[149] He M, Rutledge S, Kelly D, Palmer C, Murdoch G, Majumder N, et al. A New Genetic Disorder in Mitochondrial Fatty Acid β-Oxidation: ACAD9 Deficiency. Am J Hum Genetics 2007;81:87–103. doi:10.1086/519219 .

[150] Miyake N, Yano S, Sakai C, Hatakeyama H, Matsushima Y, Shiina M, et al. Mitochondrial Complex III Deficiency Caused by a Homozygous UQCRC2 Mutation Presenting with Neonatal-Onset Recurrent Metabolic Decompensation. Hum Mutat 2013;34:446–52. doi:10.1002/humu.22257 .

[151] Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, et al. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol 2007;3:121. doi:10.1038/msb4100155 .

[152] Orth JD, Thiele I, Palsson B. What is flux balance analysis? Nat Biotechnol 2010;28:245. doi:10.1038/nbt.1614 .

[153] Ramakrishna R, Edwards JS, McCulloch A, Palsson BO. Flux-balance analysis of mitochondrial energy metabolism: consequences of systemic stoichiometric constraints. Am J Physiology-Regulatory Integr Comp Physiology 2001;280:R695–704. doi:10.1152/ajpregu.2001.280.3.r695 .

[154] Förster J, Famili I, Fu P, Palsson B, Nielsen J. Genome-Scale Reconstruction of the Saccharomyces cerevisiae Metabolic Network. Genome Res 2003;13:244–53. doi:10.1101/gr.234503 .

[155] Hucka M, Finney A, uro H, Bolouri H, Doyle J, Kitano H, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 2003;19:524–31. doi:10.1093/bioinformatics/btg015 .

[156] Novère N, Finney A, Hucka M, Bhalla US, Campagne F, Collado-Vides J, et al. Minimum information requested in the annotation of biochemical models (MIRIAM). Nat Biotechnol 2005;23:nbt1156. doi:10.1038/nbt1156 .

[157] Swainston N, Smallbone K, Hefzi H, Dobson PD, Brewer J, Hanscho M, et al. Recon 2.2: from reconstruction to model of human metabolism. Metabolomics

2016;12:109. doi:10.1007/s11306-016-1051-4 .

[158] Thiele I, Swainston N, Fleming RM, Hoppe A, Sahoo S, Aurich MK, et al. A community-driven global reconstruction of human metabolism. Nat Biotechnol 2013;31:419. doi:10.1038/nbt.2488 .

[159] Nilsson A, Nielsen J. Genome scale metabolic modeling of cancer. Metab Eng 2017;43:103–12. doi:10.1016/j.ymben.2016.10.022 .

[160] Smith AC, Eyassu F, Mazat J-P, Robinson AJ. MitoCore: a curated constraint-based model for simulating human central metabolism. Bmc Syst Biol 2017;11:114. doi:10.1186/s12918-017-0500-7 .

[161] Boczonadi V, King MS, Smith AC, Olahova M, Bansagi B, Roos A, et al. Mitochondrial oxodicarboxylate carrier deficiency is associated with mitochondrial DNA depletion and spinal muscular atrophy–like disease. Genet Med 2018;20:1224–35. doi:10.1038/gim.2017.251 .

[162] Gaude E, Schmidt C, Gammage PA, Dugourd A, Blacker T, Chew S, et al. NADH Shuttling Couples Cytosolic Reductive Carboxylation of Glutamine with Glycolysis in Cells with Mitochondrial Dysfunction. Mol Cell 2018;69:581-593.e7. doi:10.1016/j.molcel.2018.01.034 .

[163] Heirendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, Heinken A, et al. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. Nat Protoc 2019;14:639–702. doi:10.1038/s41596-018-0098-2 .

[164] Elia M. Organ and tissue contribution to metabolic rate. In Energy Metabolism: Tissue Determinants and Cellular Corollaries, Kinney JM, Tucker HN, editors. Raven Press, New York 1992:61–80.

[165] Wang Z, Ying Z, Bosy-Westphal A, Zhang J, Schautz B, Later W, et al. Specific metabolic rates of major organs and tissues across adulthood: evaluation by mechanistic model of resting energy expenditure. Am J Clin Nutrition 2010;92:1369–

77. doi:10.3945/ajcn.2010.29885 .

[166] Wang Z, Ying Z, Bosy-Westphal A, Zhang J, Heller M, Later W, et al. Evaluation of specific metabolic rates of major organs and tissues: Comparison between men and women. Am J Hum Biol 2011;23:333–8. doi:10.1002/ajhb.21137 .

[167] Smallbone K, Simeonidis E. Flux balance analysis: A geometric perspective. J Theor Biol 2009;258:311–5. doi:10.1016/j.jtbi.2009.01.027 .

[168] Gerich JE, Meyer C, Woerle HJ, Stumvoll M. Renal Gluconeogenesis Its importance in human glucose homeostasis. Diabetes Care 2001;24:382–91. doi:10.2337/diacare.24.2.382 .

[169] Jastroch M, Divakaruni AS, Mookerjee S, Treberg JR, Brand MD. Mitochondrial proton and electron leaks. Essays Biochem 2010;47:53–67. doi:10.1042/bse0470053 .

[170] Shadel GS, Horvath TL. Mitochondrial ROS Signaling in Organismal Homeostasis. Cell 2015;163:560–9. doi:10.1016/j.cell.2015.10.001 .

[171] Ježek J, Cooper KF, Strich R. Reactive Oxygen Species and Mitochondrial Dynamics: The Yin and Yang of Mitochondrial Dysfunction and Cancer Progression. Antioxidants 2018;7:13. doi:10.3390/antiox7010013 .

[172] McCommis KS, Finck BN. Mitochondrial pyruvate transport: a historical perspective and future research directions. Biochem J 2015;466:443–54. doi:10.1042/bj20141171 .

[173] Kang PB, Hunter JV, Kaye EM. Lactic Acid Elevation in Extramitochondrial Childhood Neurodegenerative Diseases. J Child Neurol 2001;16:657–60. doi:10.1177/088307380101600906 .

[174] Meyer J. Proline transport in rat liver mitochondria. Arch Biochem Biophys 1977;178:387–95. doi:10.1016/0003-9861(77)90208-9 .

[175] Mráček T, Drahota Z, Houštěk J. The function and the role of the mitochondrial glycerol-3-phosphate dehydrogenase in mammalian tissues. Biochimica Et Biophysica Acta Bba - Bioenergetics 2013;1827:401–10. doi:10.1016/j.bbabio.2012.11.014 .

[176] Koza RA, Kozak UC, Brown LJ, Leiter EH, Macdonald MJ, Kozak LP. Sequence and Tissue-Dependent RNA Expression of Mouse FAD-Linked Glycerol-3-Phosphate Dehydrogenase. Arch Biochem Biophys 1996;336:97–104. doi:10.1006/abbi.1996.0536 .

[177] Schönfeld P, Reiser G. Why does Brain Metabolism not Favor Burning of Fatty Acids to Provide Energy? - Reflections on Disadvantages of the Use of Free Fatty Acids as Fuel for Brain. J Cereb Blood Flow Metabolism 2013;33:1493–9. doi:10.1038/jcbfm.2013.128 .

[178] Miyadera H, Shiomi K, Ui H, Yamaguchi Y, Masuma R, Tomoda H, et al. Atpenins, potent and specific inhibitors of mitochondrial complex II (succinate-ubiquinone oxidoreductase). Proc National Acad Sci 2003;100:473–7. doi:10.1073/pnas.0237315100 .

[179] Guo L, Shestov AA, Worth AJ, Nath K, Nelson DS, Leeper DB, et al. Inhibition of Mitochondrial Complex II by the Anticancer Agent Lonidamine. J Biol Chem 2016;291:42–57. doi:10.1074/jbc.m115.697516 .

[180] Pozza E, Dando I, Pacchiana R, Liboi E, Scupoli M, Donadelli M, et al. Regulation of succinate dehydrogenase and role of succinate in cancer. Semin Cell Dev Biol 2019. doi:10.1016/j.semcdb.2019.04.013 .

[181] Gnoni GV, Priore P, Geelen MJ, Siculella L. The mitochondrial citrate carrier: Metabolic role and regulation of its activity and expression. Iubmb Life 2009;61:987–94. doi:10.1002/iub.249 .

[182] Consortium T. The Universal Protein Resource (UniProt). Nucleic Acids Res

2008;36:D190–5. doi:10.1093/nar/gkm895 .

[183] Mootha VK, Bunkenborg J, Olsen JV, Hjerrild M, Wisniewski JR, Stahl E, et al. Integrated Analysis of Protein Composition, Tissue Diversity, and Gene Regulation in Mouse Mitochondria. Cell 2003;115:629–40. doi:10.1016/s0092-8674(03)00926-7 .

[184] Villeneuve LM, Stauch KL, Fox HS. Data for mitochondrial proteomic alterations in the developing rat brain. Data Brief 2014;1:42–5. doi:10.1016/j.dib.2014.07.002 .

[185] Deng W-J, Nie S, Dai J, Wu J-R, Zeng R. Proteome, Phosphoproteome, and Hydroxyproteome of Liver Mitochondria in Diabetic Rats at Early Pathogenic Stages. Mol Cell Proteomics 2010;9:100–16. doi:10.1074/mcp.m900020-mcp200 .

[186] Hodge K, Have S, Hutton L, Lamond AI. Cleaning up the masses: Exclusion lists to reduce contamination with HPLC-MS/MS. J Proteomics 2013;88:92–103. doi:10.1016/j.jprot.2013.02.023 .

[187] Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, et al. Towards a knowledge-based Human Protein Atlas. Nat Biotechnol 2010;28:1248. doi:10.1038/nbt1210-1248 .

[188] Kumar A, Agarwal S, Heyman JA, Matson S, Heidtman M, Piccirillo S, et al. Subcellular localization of the yeast proteome. Gene Dev 2002;16:707–19. doi:10.1101/gad.970902 .

[189] Omura T. Mitochondria-Targeting Sequence, a Multi-Role Sorting Sequence Recognized at All Steps of Protein Import into Mitochondria. J Biochem 1998;123:1010–6. doi:10.1093/oxfordjournals.jbchem.a022036 .

[190] Bannai H, Tamada Y, Maruyama O, Nakai K, Miyano S. Extensive feature detection of N-terminal protein sorting signals. Bioinformatics 2002;18:298–305. doi:10.1093/bioinformatics/18.2.298 .

[191] Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence. J Mol Biol 2000;300:1005–16. doi:10.1006/jmbi.2000.3903 .

[192] Claros MG, Vincens P. Computational Method to Predict Mitochondrially Imported Proteins and their Targeting Sequences. Eur J Biochem 1996;241:779–86. doi:10.1111/j.1432-1033.1996.00779.x .

[193] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems 2016.

[194] Platt JC. Probabilistic outputs for support vector machines and comparison to regularized like-lihood methods. In Advances in Large Margin Classifiers, A. Smola, P. Bartlett, B. Sch¨olkopf, and D. Schuurmans, editors. MIT Press, Cambridge, MA, 2000.

[195] Fountoulakis M, Berndt P, Langen H, Suter L. The rat liver mitochondrial proteins. Electrophoresis 2002;23:311–28. doi:10.1002/1522-2683(200202)23:2<311::aid-elps311>3.0.co;2-0 .

[196] Devreese B, Vanrobaeys F, Smet J, Beeumen J, Coster R. Mass spectrometric identification of mitochondrial oxidative phosphorylation subunits separated by two-dimensional blue-native polyacrylamide gel electrophoresis. Electrophoresis 2002;23:2525–33. doi:10.1002/1522-2683(200208)23:15<2525::aid-elps2525>3.0.co;2-i .

[197] Taylor SW, Fahy E, Zhang B, Glenn GM, Warnock DE, Wiley S, et al. Characterization of the human heart mitochondrial proteome. Nat Biotechnol 2003;21:nbt793. doi:10.1038/nbt793 .

[198] Lescuyer P, Strub J, Luche S, Diemer H, Martinez P, Dorsselaer A, et al. Progress in the definition of a reference human mitochondrial proteome. Proteomics 2003;3:157–67. doi:10.1002/pmic.200390024 .

[199] Fountoulakis M, Schlaeger E. The mitochondrial proteins of the neuroblastoma cell line IMR-32. Electrophoresis 2003;24:260–75. doi:10.1002/elps.200390022 .

[200] Cruz S, Xenarios I, Langridge J, Vilbois F, Parone PA, Martinou J-C. Proteomic Analysis of the Mouse Liver Mitochondrial Inner Membrane. J Biol Chem 2003;278:41566–71. doi:10.1074/jbc.m304940200 .

[201] Gaucher SP, Taylor SW, Fahy E, Zhang B, Warnock DE, Ghosh SS, et al. Expanded Coverage of the Human Heart Mitochondrial Proteome Using Multidimensional Liquid Chromatography Coupled with Tandem Mass Spectrometry. J Proteome Res 2004;3:495–505. doi:10.1021/pr034102a .

[202] Fukada K, Zhang F, Vien A, Cashman NR, Zhu H. Mitochondrial Proteomic Analysis of a Cell Line Model of Familial Amyotrophic Lateral Sclerosis. Mol Cell Proteomics 2004;3:1211–23. doi:10.1074/mcp.m400094-mcp200 .

[203] Jiang X-S, Dai J, Sheng Q-H, Zhang L, Xia Q-C, Wu J-R, et al. A Comparative Proteomic Strategy for Subcellular Proteome Research Icat Approach Coupled with Bioinformatics Prediction to Ascertain Rat Liver Mitochondrial Proteins and Indication of Mitochondrial Localization for Catalase. Mol Cell Proteomics 2005;4:12–34. doi:10.1074/mcp.m400079-mcp200 .

[204] Rezaul K, Wu L, Mayya V, Hwang S-I, Han D. A Systematic Characterization of Mitochondrial Proteome from Human T Leukemia Cells. Mol Cell Proteomics 2005;4:169–81. doi:10.1074/mcp.m400115-mcp200 .

[205] Xie J, Techritz S, Haebel S, Horn A, Neitzel H, Klose J, et al. A two-dimensional electrophoretic map of human mitochondrial proteins from immortalized lymphoblastoid cell lines: A prerequisite to study mitochondrial disorders in patients. Proteomics 2005;5:2981–99. doi:10.1002/pmic.200401191 .

[206] Scheffler NK, Miller SW, Carroll AK, Anderson C, Davis RE, Ghosh SS, et al. Two-dimensional electrophoresis and mass spectrometric identification of mitochondrial proteins from an SH-SY5Y neuroblastoma cell line. Mitochondrion

2001;1:161–79. doi:10.1016/s1567-7249(01)00007-1 .

[207] Forner F, Foster LJ, Campanaro S, Valle G, Mann M. Quantitative Proteomic Comparison of Rat Mitochondria from Muscle, Heart, and Liver. Mol Cell Proteomics 2006;5:608–19. doi:10.1074/mcp.m500298-mcp200 .

[208] Ruiz-Romero C, López-Armada MJ, Blanco FJ. Mitochondrial proteomic characterization of human normal articular chondrocytes. Osteoarthr Cartilage 2006;14:507–18. doi:10.1016/j.joca.2005.12.004 .

[209] Calvo S, Jain M, Xie X, Sheth SA, Chang B, Goldberger OA, et al. Systematic identification of human mitochondrial disease genes through integrative genomics. Nat Genet 2006;38:576–82. doi:10.1038/ng1776 .

[210] Kislinger T, Cox B, Kannan A, Chung C, Hu P, Ignatchenko A, et al. Global Survey of Organ and Organelle Protein Expression in Mouse: Combined Proteomic and Transcriptomic Profiling. Cell 2006;125:173–86. doi:10.1016/j.cell.2006.01.044 .

[211] Foster LJ, de Hoog CL, Zhang Y, Zhang Y, Xie X, Mootha VK, et al. A Mammalian Organelle Map by Protein Correlation Profiling. Cell 2006;125:187–99. doi:10.1016/j.cell.2006.03.022 .

[212] Reifschneider NH, Goto S, Nakamoto H, Takahashi R, Sugawa M, Dencher NA, et al. Defining the Mitochondrial Proteomes from Five Rat Organs in a Physiologically Significant Context Using 2D Blue-Native/SDS-PAGE. J Proteome Res 2006;5:1117–32. doi:10.1021/pr0504440 .

[213] McDonald T, Sheng S, Stanley B, Chen D, Ko Y, Cole RN, et al. Expanding the Subproteome of the Inner Mitochondria Using Protein Separation Technologies One- and Two-dimensional Liquid Chromatography and Two-dimensional Gel Electrophoresis. Mol Cell Proteomics 2006;5:2392–411. doi:10.1074/mcp.t500036-mcp200 .

[214] Wu L, Hwang S-I, Rezaul K, Lu LJ, Mayya V, Gerstein M, et al. Global Survey

of Human T Leukemic Cells by Integrating Proteomics and Transcriptomics Profiling. Mol Cell Proteomics 2007;6:1343–53. doi:10.1074/mcp.m700017-mcp200 .

[215] Jin J, Davis J, Zhu D, Kashima DT, Leroueil M, Pan C, et al. Identification of novel proteins affected by rotenone in mitochondria of dopaminergic cells. Bmc Neurosci 2007;8:67. doi:10.1186/1471-2202-8-67 .

[216] Lee Y, Boelsterli UA, Lin Q, Chung MC. Proteomics profiling of hepatic mitochondria in heterozygous Sod2+/− mice, an animal model of discreet mitochondrial oxidative stress. Proteomics 2008;8:555–68. doi:10.1002/pmic.200700795 .

[217] Zhang J, Li X, Mueller M, Wang Y, Zong C, Deng N, et al. Systematic characterization of the murine mitochondrial proteome using functionally validated cardiac mitochondria. Proteomics 2008;8:1564–75. doi:10.1002/pmic.200700851 .

[218] Lefort N, Yi Z, Bowen B, Glancy B, Filippis EA, Mapes R, et al. Proteome profile of functional mitochondria from human skeletal muscle using one-dimensional gel electrophoresis and HPLC-ESI-MS/MS. J Proteomics 2009;72:1046–60. doi:10.1016/j.jprot.2009.06.011 .

[219] O'Connell K, Ohlendieck K. Proteomic DIGE analysis of the mitochondria-enriched fraction from aged rat skeletal muscle. Proteomics 2009;9:5509–24. doi:10.1002/pmic.200900472 .

[220] Hadsell DL, Olea W, Wei J, Fiorotto ML, Matsunami RK, Engler DA, et al. Developmental regulation of mitochondrial biogenesis and function in the mouse mammary gland during a prolonged lactation cycle. Physiol Genomics 2010;43:271–85. doi:10.1152/physiolgenomics.00133.2010 .

[221] Egan B, Dowling P, O'Connor PL, Henry M, Meleady P, Zierath JR, et al. 2-D DIGE analysis of the mitochondrial proteome from human skeletal muscle reveals time course-dependent remodelling in response to 14 consecutive days of

endurance exercise training. Proteomics 2011;11:1413–28. doi:10.1002/pmic.201000597 .

[222] Musicco C, Capelli V, Pesce V, Timperio A, Calvani M, Mosconi L, et al. Rat liver mitochondrial proteome: Changes associated with aging and acetyl-L-carnitine treatment. J Proteomics 2011;74:2536–47. doi:10.1016/j.jprot.2011.05.041 .

[223] Chen X, Cui Z, Wei S, Hou J, Xie Z, Peng X, et al. Chronic high glucose induced INS-1β cell mitochondrial dysfunction: A comparative mitochondrial proteome with SILAC. Proteomics 2013;13:3030–9. doi:10.1002/pmic.201200448 .

[224] Villeneuve L, Tiede LM, Morsey B, Fox HS. Quantitative proteomics reveals oxygen-dependent changes in neuronal mitochondria affecting function and sensitivity to rotenone. J Proteome Res 2013;12:4599–606. doi:10.1021/pr400758d .

[225] Stauch KL, Purnell PR, Fox HS. Quantitative Proteomics of Synaptic and Nonsynaptic Mitochondria: Insights for Synaptic Mitochondrial Vulnerability. J Proteome Res 2014;13:2620–36. doi:10.1021/pr500295n .

[226] Shekari F, Nezari H, Larijani M, Han C-L, Baharvand H, Chen Y-J, et al. Proteome analysis of human embryonic stem cells organelles. J Proteomics 2017;162:108–18. doi:10.1016/j.jprot.2017.04.017 .

[227] Rabilloud T, Kieffer S, Procaccio V, Louwagie M, Courchesne PL, Patterson SD, et al. Two-dimensional electrophoresis of human placental mitochondria and protein identification by mass spectrometry: Toward a human mitochondrial proteome. Electrophoresis 1998;19:1006–14. doi:10.1002/elps.1150190616 .

[228] Hung V, Zou P, Rhee H-W, Udeshi ND, Cracan V, Svinkina T, et al. Proteomic Mapping of the Human Mitochondrial Intermembrane Space in Live Cells via Ratiometric APEX Tagging. Mol Cell 2014;55:332–41. doi:10.1016/j.molcel.2014.06.003 .

[229] Breker M, Gymrek M, Schuldiner M. A novel single-cell screening platform

reveals proteome plasticity during yeast stress responses. J Cell Biology 2013;200:839–50. doi:10.1083/jcb.201301120 .

[230] Tkach JM, Yimit A, Lee AY, Riffle M, Costanzo M, Jaschob D, et al. Dissecting DNA damage response pathways by analysing protein localization and abundance changes during DNA replication stress. Nat Cell Biol 2012;14:966. doi:10.1038/ncb2549 .

[231] Matsuyama A, Arai R, Yashiroda Y, Shirai A, Kamata A, Sekido S, et al. ORFeome cloning and global analysis of protein localization in the fission yeast Schizosaccharomyces pombe. Nat Biotechnol 2006;24:841–7. doi:10.1038/nbt1222 .

[232] Ozawa T, Sako Y, Sato M, Kitamura T, Umezawa Y. A genetic approach to identifying mitochondrial proteins. Nat Biotechnol 2003;21:nbt791. doi:10.1038/nbt791 .

[233] Huh W-K, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, et al. Global analysis of protein localization in budding yeast. Nature 2003;425:686. doi:10.1038/nature02026 .

[234] Gao J, Schatton D, Martinelli P, Hansen H, Pla-Martin D, Barth E, et al. CLUH regulates mitochondrial biogenesis by binding mRNAs of nuclear-encoded mitochondrial proteins. J Cell Biology 2014;207:213–23. doi:10.1083/jcb.201403129 .

[235] Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 2009;19:327–35. doi:10.1101/gr.073585.107 .

[236] Karnkowska A, Vacek V, Zubáčová Z, Treitli SC, Petrželková R, Eme L, et al. A Eukaryote without a Mitochondrial Organelle. Curr Biol 2016;26:1274–84. doi:10.1016/j.cub.2016.03.053 .

[237] Wang J, Chen Q, Chen Y. Advances in Neural Networks – ISNN 2004,

International Symposium on Neural Networks, Dalian, China, August 2004, Proceedings, Part I 2004:512–7. doi:10.1007/978-3-540-28647-9_85 .

[238] Binder JX, Pletscher-Frankild S, Tsafou K, Stolte C, O'Donoghue SI, Schneider R, et al. COMPARTMENTS: unification and visualization of protein subcellular localization evidence. Database 2014;2014:bau012. doi:10.1093/database/bau012 .

[239] Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res 2017;45:D362–8. doi:10.1093/nar/gkw937 .

[240] Tarca AL, Romero R, Draghici S. Analysis of microarray experiments of gene expression profiling. Am J Obstet Gynecol 2006;195:373–88. doi:10.1016/j.ajog.2006.07.001 .

[241] Jan CH, Williams CC, Weissman JS. Principles of ER cotranslational translocation revealed by proximity-specific ribosome profiling. Science 2014;346:1257521. doi:10.1126/science.1257521 .

[242] van der Maaten L, Hinton G. Visualizing Data using t-SNE. 2008.

[243] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System 2016:785–94. doi:10.1145/2939672.2939785 .

[244] Dimmock D, Maranda B, Dionisi-Vici C, Wang J, Kleppe S, Fiermonte G, et al. Citrin deficiency, a perplexing global disorder. Mol Genet Metab 2009;96:44–9. doi:10.1016/j.ymgme.2008.10.007 .

[245] Sato M, Sato K. Maternal inheritance of mitochondrial DNA by diverse mechanisms to eliminate paternal mitochondrial DNA. Biochimica Et Biophysica Acta Bba - Mol Cell Res 2013;1833:1979–84. doi:10.1016/j.bbamcr.2013.03.010 .

[246] Tuppen H, Blakely EL, Turnbull DM, Taylor RW. Mitochondrial DNA mutations

and human disease. Biochimica Et Biophysica Acta Bba - Bioenergetics 2010;1797:113–28. doi:10.1016/j.bbabio.2009.09.005 .

[247] Sallevelt SC, de Die-Smulders CE, Hendrickx AT, Hellebrekers DM, de Coo IF, Alston CL, et al. De novo mtDNA point mutations are common and have a low recurrence risk. J Med Genet 2017;54:73. doi:10.1136/jmedgenet-2016-103876 .

[248] Wallace DC, Chalkia D. Mitochondrial DNA Genetics and the Heteroplasmy Conundrum in Evolution and Disease. Csh Perspect Biol 2013;5:a021220. doi:10.1101/cshperspect.a021220 .

[249] Gorman GS, Schaefer AM, Ng Y, Gomez N, Blakely EL, Alston CL, et al. Prevalence of nuclear and mitochondrial DNA mutations related to adult mitochondrial disease. Ann Neurol 2015;77:753–9. doi:10.1002/ana.24362 .

[250] Alston CL, Rocha MC, Lax NZ, Turnbull DM, Taylor RW. The genetics and pathology of mitochondrial disease. J Pathology 2017;241:236–50. doi:10.1002/path.4809 .

[251] Angelini C, Bello L, inazzi, Ferrati C. Mitochondrial disorders of the nuclear genome. Acta Myologica Myopathies Cardiomyopathies Official J Mediterr Soc Myology Ed Gaetano Conte Acad Study Striated Muscle Dis 2009;28:16–23.

[252] Tang S, Wang J, Zhang V, Li F, Landsverk M, Cui H, et al. Transition to Next Generation Analysis of the Whole Mitochondrial Genome: A Summary of Molecular Defects. Hum Mutat 2013;34:882–93. doi:10.1002/humu.22307 .

[253] Alston CL, Compton AG, Formosa LE, Strecker V, Oláhová M, Haack TB, et al. Biallelic Mutations in TMEM126B Cause Severe Complex I Deficiency with a Variable Clinical Phenotype. Am J Hum Genetics 2016;99:217–27. doi:10.1016/j.ajhg.2016.05.021 .

[254] Haack TB, Danhauser K, Haberberger B, Hoser J, Strecker V, Boehm D, et al. Exome sequencing identifies ACAD9 mutations as a cause of complex I deficiency.

Nat Genet 2010;42:1131. doi:10.1038/ng.706 .

[255] Hartmannová H, Piherová L, Tauchmannová K, Kidd K, Acott PD, Crocker JF, et al. Acadian variant of Fanconi syndrome is caused by mitochondrial respiratory chain complex I deficiency due to a non-coding mutation in complex I assembly factor NDUFAF6. Hum Mol Genet 2016;25:4062–79. doi:10.1093/hmg/ddw245 .

[256] Theunissen TE, Nguyen M, Kamps R, Hendrickx AT, Sallevelt SC, Gottschalk RW, et al. Whole Exome Sequencing Is the Preferred Strategy to Identify the Genetic Defect in Patients With a Probable or Possible Mitochondrial Cause. Frontiers Genetics 2018;9:400. doi:10.3389/fgene.2018.00400 .

[257] Pronicka E, Piekutowska-Abramczuk D, Ciara E, Trubicka J, Rokicki D, Karkucińska-Więckowska A, et al. New perspective in diagnostics of mitochondrial disorders: two years' experience with whole-exome sequencing at a national paediatric centre. J Transl Med 2016;14:174. doi:10.1186/s12967-016-0930-9 .

[258] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297–303. doi:10.1101/gr.107524.110 .

[259] Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. Science 2012;337:64–9. doi:10.1126/science.1219240 .

[260] Nelson MR, Wegmann D, Ehm MG, Kessner D, Jean P, Verzilli C, et al. An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. Science 2012;337:100–4. doi:10.1126/science.1217876 .

[261] Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 2005;33:D514–7. doi:10.1093/nar/gki033 .

[262] Mitochondrial disorders. Neuromuscular disease centre, Washington University, St. Louis, MO, USA. https://neuromuscular.wustl.edu/mitosyn.html .

[263] Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res 2017;45:D840–5. doi:10.1093/nar/gkw971 .

[264] Taylor RW, Pyle A, Griffin H, Blakely EL, Duff J, He L, et al. Use of Whole-Exome Sequencing to Determine the Genetic Basis of Multiple Mitochondrial Respiratory Chain Complex Deficiencies. Jama 2014;312:68–77. doi:10.1001/jama.2014.7184 .

[265] Calvo SE, Compton AG, Hershman SG, Lim S, Lieber DS, Tucker EJ, et al. Molecular Diagnosis of Infantile Mitochondrial Disease with Targeted Next-Generation Sequencing. Sci Transl Med 2012;4:118ra10-118ra10. doi:10.1126/scitranslmed.3003310 .

[266] Williamson R, Kessling AM. Ciba Foundation Symposium 149 - Human Genetic Information: Science, Law and Ethics 2007:63–80. doi:10.1002/9780470513903.ch6 .

[267] Moreau Y, Tranchevent L-C. Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nat Rev Genet 2012;13:523. doi:10.1038/nrg3253 .

[268] Sevimoglu T, Arga K. The role of protein interaction networks in systems biomedicine. Comput Struct Biotechnology J 2014;11:22–7. doi:10.1016/j.csbj.2014.08.008 .

[269] Wang PI, Marcotte EM. It’s the machine that matters: Predicting gene function and phenotype from protein networks. J Proteomics 2010;73:2277–89. doi:10.1016/j.jprot.2010.07.005 .

[270] Jordán F, Nguyen T-P, Liu W. Studying protein–protein interaction networks: a

systems view on diseases. Brief Funct Genomics 2012;11:497–504. doi:10.1093/bfgp/els035 .

[271] Brun C, Herrmann C, Guénoche A. Clustering proteins from interaction networks for the prediction of cellular functions. Bmc Bioinformatics 2004;5:95. doi:10.1186/1471-2105-5-95 .

[272] Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet 2011;12:56. doi:10.1038/nrg2918 .

[273] Oti M, Snel B, Huynen M, Brunner H. Predicting disease genes using protein–protein interactions. J Med Genet 2006;43:691. doi:10.1136/jmg.2006.041376 .

[274] Suratanee A, Plaimas K. Network-based association analysis to infer new disease-gene relationships using large-scale protein interactions. Plos One 2018;13:e0199435. doi:10.1371/journal.pone.0199435 .

[275] Xu J, Li Y. Discovering disease-genes by topological features in human protein–protein interaction network. Bioinformatics 2006;22:2800–5. doi:10.1093/bioinformatics/btl467 .

[276] Rusu V, Hoch E, Mercader JM, Tenen DE, Gymrek M, Hartigan CR, et al. Type 2 Diabetes Variants Disrupt Function of SLC16A11 through Two Distinct Mechanisms. Cell 2017;170:199-212.e20. doi:10.1016/j.cell.2017.06.011 .

[277] Lage K, Hansen N, Karlberg OE, Eklund AC, Roque FS, Donahoe PK, et al. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. Proc National Acad Sci 2008;105:20870–5. doi:10.1073/pnas.0810772105 .

[278] Feiglin A, Allen BK, Kohane IS, Kong S. Comprehensive Analysis of Tissue-wide Gene Expression and Phenotype Data Reveals Tissues Affected in Rare Genetic Disorders. Cell Syst 2017;5:140-148.e2. doi:10.1016/j.cels.2017.06.016 .

[279] Smith CL. Comparative phylogenetic exploration of the human mitochondrial proteome: Insights into disease and metabolism 2019. doi:10.17863/cam.31654 .

[280] Ho D, Schierding W, Wake M, Saffery R, O'Sullivan J. Machine Learning SNP Based Prediction for Precision Medicine. Frontiers Genetics 2019;10:267. doi:10.3389/fgene.2019.00267 .

[281] Asif M, Martiniano HF, Vicente AM, Couto FM. Identifying disease genes using machine learning and gene functional similarities, assessed through Gene Ontology. Plos One 2018;13:e0208626. doi:10.1371/journal.pone.0208626 .

[282] Peng J, Guan J, Shang X. Predicting Parkinson's Disease Genes Based on Node2vec and Autoencoder. Frontiers Genetics 2019;10:226. doi:10.3389/fgene.2019.00226 .

[283] Liao Z, Li D, Wang X, Li L, Zou Q. Cancer Diagnosis Through IsomiR Expression with Machine Learning. Method Current Bioinformatics 2018;13:57-63. doi:10.2174/1574893611666160609081155 .

[284] Aiyar R, Gagneur J, Steinmetz L. Identification of mitochondrial disease genes through integrative analysis of multiple datasets. Methods 2008;46:248-255. doi:10.1016/j.ymeth.2008.10.002 .

[285] Krogh A. What are artificial neural networks?. Nature Biotechnology 2008;2: nbt1386. doi:10.1038/nbt1386 .

[286] Luo P, Li Y, Tian L, Wu F. Enhancing the prediction of disease-gene associations with multimodal deep learning. Bioinformatics (Oxford, England) 2019. doi:10.1093/bioinformatics/btz155 .

[287] Sulem P, Helgason H, Oddson A, Stefansson H, Gudjonsson S, Zink F, et.al. Identification of a large set of rare complete human knockouts. Nature Genetics 2015;47:ng.3243. doi:10.1038/ng.3243 .

[288] Saleheen D, Natarajan P, Armean I, Zhao W, Rasheed A, Khetarpal S, et.al. Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. Nature 2017;544-235. doi:10.1038/nature22034 .

[289] Narasimhan V, Hunt K, Mason D, Baker C, Karczewski K, Barnes M, et.al. Health and population effects of rare gene knockouts in adult humans with related parents. Science 2016;352:474-477. doi:10.1126/science,aac8624 .

[290] Li T, Wernersson R, Hansen R, Horn H, Mercer J, Slodkowicz G, et.al. A scored human protein–protein interaction network to catalyze genomic interpretation. Nature Methods 2016;14:4083. doi:10.1038/nmeth.4083 .

[291] Kamburov A, Pentchev K, Galicka H, Wierling C, Lehrach H, Herwig R. ConsensusPathDB: toward a more complete picture of cell biology. Nucleic Acids Research 2011;39. doi:10.1093/nar/gkq1156 .

[292] Calderone A, Castagnoli L, Cesareni G. mentha: a resource for browsing integrated protein-interaction networks. Nature Methods 2013;10:2561. doi:10.1038/nmeth.2561 .

[293] Kotlyar M, Pastrello C, Malik Z, Jurisica I. IID 2018 update: context-specific physical protein–protein interactions in human model organisms and domesticated species. Nucleic Acids Research 2018;47. doi:10.1093/nar/gky1037 .

[294] Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal Complex Systems. 2006. http://igraph.org .

[295] Travençolo B, Costa L. Accessibility in complex networks. Physics Letters A 2008;373:89-95. doi:10.1016/j.physleta.2008.10,069 .

[296] Becker E, Robisson B, Chapple C, Guénoche A, Brun C. Multifunctional proteins revealed by overlapping clustering in protein interaction network. Bioinformatics 2012;28:84-90. doi:10.1093/bioinformatics/btr621 .

[297] Parodi-Rullán R, Chapa-Dubocq X, Javadov S. Acetylation of Mitochondrial Proteins in the Heart: The Role of SIRT3. Frontiers in Physiology 2018;9:1094. doi:10.3389/fphys.2018.01094 .

[298] Baeza J, Smallegan M, Denu J. Mechanisms and Dynamics of Protein Acetylation in Mitochondria. Trends in Biochemical Sciences 2016;41:231-244. doi:10.1016/j.tibs,2015.12.006 .

[299] Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. bioRxiv 2016. doi:10.1101/030338 .

[300] Wilcoxon F. Individual Comparisons by Ranking Methods. Biometrics Bulletin 1945;1:80. doi:10.2307/3001968 .

[301] Ruder S. An overview of gradient descent optimization algorithms. 2016.

[302] Kingma D, Ba J. Adam: A Method for Stochastic Optimization, International Conference on Learning Representations. 2014.

[303] Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. 2015.

[304] Ioffe S. Batch Renormalization: Towards Reducing Minibatch Dependence in Batch-Normalized Models. 2017.

[305] Nair V, Hinton G. Rectified Linear Units Improve Restricted Boltzmann Machines, In Proceedings of ICML 2010;27:807-814 .

[306] Maas AL, Hannun AY, Ng AY. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In Proceedings of ICML 2013;30.

[307] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research 2014.

[308] He K, Zhang X, Ren S, Sun J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. IEEE International Conference on Computer Vision (ICCV) 2015. doi:10.1109/iccv.2015.123 .

[309] Drummond D, Bloom J, Adami C, Wilke C, Arnold F. Why highly expressed proteins evolve slowly. In Proceedings of the National Academy of Sciences of the United States of America 2005;102:14338-14343. doi:10.1073/pnas.0504070102 .

[310] Warringer J, Blomberg A. Evolutionary constraints on yeast protein size. BMC Evolutionary Biology 2006;6:61. doi:10.1186/1471-2148-6-61 .

[311] Huang D, Sherman B, Lempicki R. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Research 2009;37:1-13. doi:10.1093/nar/gkn923 .

[312] Ikeda F, Yoshida K, Toki T, Uechi T, Ishida S, Nakajima Y, et.al. Exome sequencing identified RPS15A as a novel causative gene for Diamond-Blackfan anemia. Haematologica 2016;102:e93-e96. doi:10.3324/haematol.2016.153932 .

[313] Rivière J, Bon B, Hoischen A, Kholmanskikh S, O'Roak B, Gilissen C, et.al. De novo mutations in the actin genes ACTB and ACTG1 cause Baraitser-Winter syndrome. Nature Genetics 2012;44:440. doi:10.1038/ng.1091 .

[314] Popov I, Hiatt S, Whalen S, Keren B, Ruivenkamp C, Haeringen A, et.al. A YWHAZ Variant Associated With Cardiofaciocutaneous Syndrome Activates the RAF-ERK Pathway. Frontiers in Physiology 2019;10:388. doi:10.3389/fphys.2019.00388 .

[315] Petrovski S, Küry S, Myers C, Anyane-Yeboa K, Cogné B, Bialer, et.al. Germline De Novo Mutations in GNB1 Cause Severe Neurodevelopmental Disability Hypotonia and Seizures. The American Journal of Human Genetics 2016;98:1001-1010. doi:10.1016/j.ajhg.2016.03.011 .

[316] Sacconi S, Féasson L, Antoine J, Pécheux C, Bernard R, Cobo A, et.al. A novel CRYAB mutation resulting in multisystemic disease. Neuromuscular Disorders 2012;22:66-72. doi:10.1016/j.nmd.2011.07.004 .

[317] Carpten J, Faber A, Horn C, Donoho G, Briggs S, Robbins C, et.al. A transforming mutation in the pleckstrin homology domain of AKT1 in cancer. Nature 2007;448:439. doi:10.1038/nature05933 .

[318] Irby R, Mao W, Coppola D, Kang J, Loubeau J, Trudeau W, et.al. Activating SRC mutation in a subset of advanced human colon cancers. Nature Genetics 1999;21:187-190. doi:10.1038/5971 .

[319] Magen D, Georgopoulos C, Bross P, Ang D, Segev Y, Goldsher D, et.al. Mitochondrial Hsp60 Chaperonopathy Causes an Autosomal-Recessive Neurodegenerative Disorder Linked to Brain Hypomyelination and Leukodystrophy. The American Journal of Human Genetics 2008;83:30-42. doi:10.1016/j.ajhg.2008.05.016 .

[320] Bugiardini E, Mitchell A, Rosa I, Horning-Do H, Pitmann A, Poole O, et.al. MRPS25 mutations impair mitochondrial translation and cause encephalomyopathy Human Molecular Genetics 2019;28:2711-2719. doi: 10.1093/hmg/ddz093 .

Appendices

Appendix I: The significantly different metabolites during mitochondrial dysfunction

Table 1. Metabolites identified as significantly different in the plasma of the low dose rats after 2 hours.

	Low dose vs Control						High dose vs Control			Exposure correlation
Metabolite	q-value	VIP Score	RFECV mean	Fold-change (log2)	Cons. level		q-value	Fold-change (log2)		Rho
methionine	0.012281	4.716701	1	-0.31774	9		0.053586	-0.35713		-0.36364
phenyllactate (PLA)	0.012281	4.511833	1	-1.17226	9		0.173225	-1.31542		-0.24545
2-piperidinone	0.016477	4.480894	1	-0.78807	9		0.074427	-0.79585		-0.00909
myo-inositol	0.039745	4.123371	1.142857	-0.63313	9		0.074427	-0.71509		-0.09091
proline	0.016091	4.223782	1.285714	-0.43809	9		0.029197	-0.5357		-0.37273
hypotaurine	0.038313	4.081483	1.714286	-0.71258	9		0.271484	-0.48875		0.209091
3-hydroxybutyrate (BHBA)	0.019786	3.867431	3.714286	1.551293	9		0.275634	1.616223		-0.21818
adenine	0.06531	3.638111	6.285714	-0.24584	9		0.271492	-0.3231		-0.17273
methyl glucopyranoside (alpha + beta)	0.016091	4.196769	17	-0.67324	9		0.051801	-0.86769		-0.29091
2-arachidonoylglycerophosphoinositol	0.101435	4.047362	2.285714	0.75615	8		0.41663	0.884		0.474891
arachidate (20:0)	0.071646	3.425517	11.42857	-0.29474	8		0.053586	-0.44116		-0.48182
N-acetylmethionine	0.06531	3.434251	19	-0.48507	8		0.514613	-0.29037		0.563636
palmitoleate (16:1n7)	0.016091	3.672584	27	1.260022	8		0.22228	1.692		0.218182
16-hydroxypalmitate	0.032886	3.539416	49	1.053061	8		0.053586	1.294285		0.218182
EDTA	0.131449	2.82429	4.571429	-0.2438	7		0.075714	-0.42572		-0.36364
X - 22145	0.038098	2.355309	5.428571	-1.55032	7		0.559966	-0.26621		-0.06364
X - 12818	0.016091	2.332899	7.142857	1.762761	7		0.235102	-1.12375		0.287017
1-arachidonoyl-GPC (20:4)	0.101435	3.28487	8	0.191285	7		0.198426	0.300106		0.318182

5-methyluridine (ribothymidine)	0.104472	3.104708	9.714286	-0.24992	7		0.271492	-0.26856		0.2
X - 21807	0.289481	3.643452	18	-0.69466	7		0.33287	-1.48427		0.063636
X - 12170	0.140817	3.18102	22	-0.2395	7		0.433223	-0.46481		0.272727
N-acetylmethionine sulfoxide	0.075654	3.312568	30	-1.51633	7		0.910612	0.352762		0.692485
docosapentaenoate (DPA; 22:5n3)	0.082798	2.899834	34	0.533395	7		0.151661	0.505238		-0.37273
gamma-glutamylalanine	0.090656	3.402096	38	-0.49516	7		0.198426	-0.55357		0.072727
mannose	0.071646	3.04353	42	0.480191	7		0.235102	0.492131		0.036364
palmitate (16:0)	0.06531	3.36832	61	0.421961	7		0.344006	0.498867		0.027273
acetoacetate	0.026314	3.38961	70	1.593893	7		0.271492	1.225847		-0.39091
linolenate (18:3n3 or 3n6)	0.059256	3.051576	93	0.840903	7		0.282601	0.940057		-0.10909
N-6-trimethyllysine	0.157945	2.55609	21	-0.23009	6		0.095199	-0.40218		-0.7
glycylleucine	0.136622	2.665973	39	-0.44882	6		0.235102	-0.6707		-0.5
2-hydroxyoctanoate	0.149108	2.532576	46	-0.29504	6		0.235102	-0.54781		-0.41818
3-hydroxyisobutyrate	0.104472	3.197991	47	0.613542	6		0.173225	0.782866		0.163636
cyclo(L-phe-L-pro)	0.136622	2.626884	50	-0.46785	6		0.235102	-0.72262		-0.54545
mead acid (20:3n9)	0.101435	3.282024	55	0.622016	6		0.282601	0.90717		-0.01818
cyclo(leu-pro)	0.130397	2.954024	57	-0.65216	6		0.25297	-0.92305		-0.25455
gamma-glutamylmethionine	0.115773	2.717151	64	-0.35245	6		0.271492	-0.23872		0.2
pentadecanoate (15:0)	0.101435	2.917625	78	0.232983	6		0.423508	0.242615		-0.19091
methionine sulfoxide	0.127937	2.586786	83	-0.53665	6		0.146096	-0.70692		-0.44545
alanine	0.100223	3.082259	87	-0.32146	6		0.235102	-0.42443		-0.03636
4-guanidinobutanoate	0.149108	2.659812	88	-0.5557	6		0.235102	-0.80016		-0.19091
mevalonate	0.10294	2.613088	95	-0.71877	6		0.151661	-0.78286		-0.26424
X - 21477	0.130397	3.003553	98	-0.43229	6		0.238267	-0.7229		-0.29091
linoleoylcarnitine	0.042254	3.424846	105	0.913708	6		0.306171	0.794623		-0.23636
linoleate (18:2n6)	0.071646	3.236731	109	0.339975	6		0.410737	0.409442		-0.25455
X - 17761	0.080138	2.739125	121	-1.70743	6		0.235102	1.390799		0.045455

10-heptadecenoate (17:1n7)	0.06531	3.043511	140	0.771307	6		0.25297	0.909293		0.009091
oleoylcarnitine (C18)	0.045427	3.381146	146	1.128206	6		0.25297	1.209401		-0.14545
myristoleate (14:1n5)	0.082055	2.945635	153	0.934009	6		0.254643	1.247202		-0.01818
1-methylimidazoleacetate	0.224863	2.41335	15	-0.57356	5		0.306171	-0.73634		0.127273
2-arachidonoyl-GPC (20:4)	0.230372	2.196027	20	0.22863	5		0.235102	0.470731		0.127273
12,13-DiHOME	0.196143	2.443693	23	0.658725	5		0.888967	0.097586		-0.33636
citrate	0.180933	2.101092	25	0.189108	5		0.266965	0.332497		0.309091
valylglycine	0.162963	2.806258	33	-0.75067	5		0.874827	-0.05203		0.454545
cinnamate	0.157945	2.559908	45	0.620076	5		0.100037	0.84788		0.223235
gamma-glutamylleucine	0.157945	2.553039	58	-0.19642	5		0.275634	-0.19415		0.118182
X - 12411	0.136622	2.205701	65	-1.27729	5		0.514613	-0.48698		0.236364
X - 12442	0.275997	3.429089	82	-0.27544	5		0.434135	0.912886		-0.29091
6-oxopiperidine-2-carboxylic acid	0.155852	2.583731	91	-0.51494	5		0.198426	-0.54285		0.018182
margarate (17:0)	0.101435	3.074673	107	0.443718	5		0.474127	0.403184		-0.17273
dihomolinolenate (20:3n3 or 3n6)	0.110011	2.542434	136	0.338591	5		0.235102	0.633932		0.363636
prolylphenylalanine	0.116279	2.733117	174	-0.87806	5		0.238267	-0.91074		-0.08182
hydroxyproline	0.149108	2.55478	184.4286	-0.20354	5		0.077761	-0.30195		-0.12727
palmitoylcarnitine (C16)	0.080138	2.782585	222	0.819122	5		0.344006	0.856727		-0.19091
3-phosphoglycerate	0.17106	2.203112	26	-1.04395	4		0.433223	-0.98014		0.318907
X - 11612	0.192695	2.202644	29	-0.98315	4		0.598962	0.886768		-0.00909
X - 20587	0.230372	2.106959	32	-0.43719	4		0.476894	-0.3417		0.145455
taurine	0.216862	2.127304	48	-0.43243	4		0.39005	-0.25295		0.163636
2-hydroxydecanoate	0.260669	2.073432	51	-0.50993	4		0.876376	-0.05963		0.390909
O-sulfo-L-tyrosine	0.180933	2.25161	52	-0.20709	4		0.246507	-0.29772		-0.26364
trans-urocanate	0.230372	2.079909	54	-0.9032	4		0.235102	-1.0281		-0.19091
4-hydroxyphenylpyruvate	0.207807	2.169578	67	-0.44139	4		0.33287	-0.44518		-0.05455
valerylcarnitine (C5)	0.157945	2.206882	77	0.5905	4		0.649001	0.256886		-0.60909

glutaroylcarnitine (C5)	0.192695	2.239182	81	-0.4102	4		0.306171	-0.83811		-0.15455
carnitine	0.213734	2.097068	84	-0.2123	4		0.383026	-0.23868		0.190909
guanidinoacetate	0.140817	2.486186	119	-0.82213	4		0.236643	-0.74797		0.327273
4-hydroxyphenylacetyl glycine	0.115773	2.245124	124	-1.2259	4		0.235102	-1.28016		0.281818
2-hydroxybutyrate (AHB)	0.101098	2.471234	171	1.915204	4		0.238267	2.41726		0.063636
myristoleoylcarnitine	0.115773	2.846017	232	1.017575	4		0.235102	0.689123		-0.37273
propionylcarnitine (C3)	0.172998	2.351591	104	-0.42765	3		0.306171	-0.43193		0.027273
imidazole lactate	0.235576	2.184238	111	-0.43323	3		0.250385	-0.75315		-0.24545
isoleucylglycine	0.197825	2.358543	112	-0.34749	3		0.714922	-0.14921		0.136364
isovalerylcarnitine (C5)	0.180933	2.324549	115	-0.42579	3		0.334425	-0.64347		-0.01818
valylisoleucine	0.230372	2.293493	131	-1.06605	3		0.588406	-0.47578		0.428247
N6-acetyllysine	0.202106	2.116863	133	-0.21729	3		0.235102	-0.23835		0.045455
acetylcarnitine (C2)	0.230372	2.149902	138	0.229091	3		0.832623	0.112002		-0.42727
1-stearoyl-GPE (18:0)	0.212485	2.073061	148	0.473404	3		0.25297	0.422346		-0.1
X - 19561	0.250507	2.028295	152	-0.73699	3		0.787498	-0.2271		0.646927
arabitol	0.162963	2.166538	157	-0.47475	3		0.44081	-0.24367		0.627273
erythritol	0.197825	2.245262	161	-0.34628	3		0.235102	-0.30651		0.054545
guanosine-3',5'-cyclic monophosphate (cGMP)	0.17106	2.022258	170	-0.86335	3		0.476894	-0.58956		0.318907
3-hydroxyoctanoate	0.147763	2.208096	202	0.50055	3		0.283204	0.645093		0.181818
X - 16975	0.230086	3.18324	285	0.451842	3		0.151661	-1.75537		-0.46364
myristoylcarnitine	0.130397	2.351944	303	0.862659	3		0.382218	0.893767		-0.1
X - 21329	0.242788	2.667255	507	4.202287	3		0.238267	0.438453		0.172727
X - 12730	0.964701	3.747455	526	-0.14115	3		0.074427	-1.92158		0.172727
X - 12199	0.596603	4.059631	600	0.165925	3		0.124517	-1.26713		-0.49091
X - 12721	0.649051	3.681763	684	0.370839	3		0.323973	2.109984		0.072727
docosapentaenoate (n6 DPA; 22:5n6)	0.360323	2.048139	118	0.380696	2		0.238267	0.703224		0.527273
X - 12007	0.824183	2.08026	179	0.789352	2		0.271492	-1.26897		0.090909

X - 18913	0.64621	2.429116	195	-0.16541	2		0.965822	0.136917		0.272727
S-methylcysteine	0.180933	2.279775	210	0.332549	2		0.271492	0.246003		-0.31818
alpha-CEHC	0.252738	2.070855	217	0.903424	2		0.235102	0.83757		0.081818
10-nonadecenoate (19:1n9)	0.18375	2.371745	235	0.453766	2		0.39005	0.603011		-0.04545
bilirubin	0.244614	2.072742	248	0.70409	2		0.198426	1.222308		0.182233
3-hydroxy-3-methylglutarate	0.162963	2.141036	260	0.376772	2		0.665446	0.322947		-0.37273
X - 12119	0.595042	2.631418	298	0.419286	2		0.679301	0.133615		-0.31818
stearoylcarnitine (C18)	0.18375	2.231192	305	0.371406	2		0.543047	0.206302		-0.41818
oleate (18:1n9)	0.157945	2.306246	334	0.72082	2		0.264476	1.019691		0.118182
X - 17717	0.539094	3.392825	340	-0.32665	2		0.383026	-0.85311		-0.19091
3-methyl-2-oxovalerate	0.191815	2.224986	350	0.750063	2		0.077761	0.813277		-0.07273
3-methyl-2-oxobutyrate	0.197825	2.196639	416	0.434204	2		0.202999	0.467792		-0.08182
X - 21431	0.53788	2.767055	504	0.22514	2		0.319159	-0.3707		0.354545
X - 17735	0.781855	2.893461	506	0.235801	2		0.28641	0.893997		0.090909
X - 21788	0.846795	2.72092	617	-0.14051	2		0.514613	-0.24386		0.345455
X - 12267	0.781855	3.387627	629	0.169936	2		0.235102	-1.49381		-0.21868
X - 14374	0.89537	2.665028	678	0.066252	2		0.433223	-0.26143		0.3
X - 21792	0.964701	2.500299	710	-0.00659	2		0.44081	1.079978		-0.43636
X - 12100	0.942872	2.63647	735	-0.04835	2		0.074427	-0.42887		-0.56364
X - 15497	0.874481	2.6279	736	-0.20365	2		0.37191	-0.55118		-0.04545
X - 14473	0.53788	2.085146	213	-0.24891	1		0.282601	-0.48934		-0.31818
X - 18889	0.497037	2.028188	329	0.400841	1		0.44081	-0.45459		0.218182
X - 11442	0.846795	2.370804	368	-0.10371	1		0.659974	-0.32787		0.419135
X - 12450	0.696613	2.350147	390	-0.13254	1		0.36037	0.731951		-0.22727
X - 21664	0.497037	2.442846	403	-0.10288	1		0.244843	-0.58654		-0.39091
X - 13846	0.395505	2.373643	431	0.456808	1		0.514613	-0.83665		0.364466
X - 12329	0.53788	2.191707	466	0.314165	1		0.269616	-1.68029		-0.31818
X - 17367	0.874481	2.21205	486	0.034349	1		0.832392	0.288984		0.728931
X - 22776	0.649051	2.19151	539	-0.50085	1		0.251986	-0.71276		0.336364

X - 13529	0.516351	2.417909	568	0.428657	1		0.491725	-0.265		0.236364
X - 21892	0.974914	2.301351	730	0.046361	1		0.383026	0.384672		0.127273

Table 2. Metabolites identified as significantly different in the plasma of the low dose rats after 4 hours.

	Low dose vs Control						High dose vs Control			Exposure correlation
Metabolites	**q-value**	**VIP Score**	**RFECV mean**	**Fold-change (log2)**	**Cons. level**		**q-value**	**Fold-change (log2)**		**Rho**
mevalonate	0.008585	4.830892	1	-1.02308	9		0.016239	-1.52559		-0.73636
phenyllactate (PLA)	0.005177	4.675277	1	-1.39605	9		0.275586	-0.74199		0.445455
prolylphenylalanine	0.008585	4.50887	1.571429	-0.91394	9		0.563829	-0.26221		0.445455
proline	0.008585	4.578103	2.285714	-0.79517	9		0.331937	-0.45932		-0.14545
methionine	0.027058	4.26247	3.714286	-0.53796	9		0.240253	-0.33809		-0.01818
2-piperidinone	0.038463	4.100551	4.428571	-0.98571	9		0.038997	-0.71342		0.490909
methyl glucopyranoside (alpha + beta)	0.008585	4.331417	5.142857	-0.93064	9		0.001606	-1.8482		-0.83636
5-oxoproline	0.038945	3.768579	5.857143	-0.36824	9		0.050038	-0.46381		-0.37273
6-oxopiperidine-2-carboxylic acid	0.01353	4.418407	8.285714	-0.74648	9		0.143395	-0.42963		0.063636
asparagine	0.038463	3.571622	10.85714	-0.6047	9		0.278137	-0.33996		0.127273
citrulline	0.045711	3.77694	12.57143	-0.42369	9		0.445176	-0.30306		-0.33636
alanine	0.038463	3.631242	16	-0.66264	9		0.847245	0.401375		0.272727
cyclo(L-phe-L-pro)	0.045711	3.516566	16.85714	-0.64695	9		0.019255	-0.98211		-0.65455
methionine sulfoxide	0.038463	3.953058	18.57143	-0.89021	9		0.096353	-0.94381		-0.35455
3-hydroxyisobutyrate	0.014541	3.929132	21.14286	0.646173	9		0.11064	1.621229		0.918182
cyclo(gly-pro)	0.038463	3.978582	23.71429	-0.66893	9		0.121698	-1.06576		-0.54545
carnitine	0.051665	3.331363	6.571429	-0.3006	8		0.110997	-0.46018		-0.52727
palmitoylcarnitine (C16)	0.038463	3.166767	10	1.128396	8		0.001606	1.342882		0.118182
2-aminophenol sulfate	0.051665	3.059837	19.42857	-0.78798	8		0.744624	0.23955		0.781818
gamma-glutamylmethionine	0.045711	3.456907	20.28571	-0.69225	8		0.334642	-0.43538		-0.18182
citrate	0.096009	2.663198	24.71429	0.239368	8		0.180597	0.670165		0.427273

gamma-glutamylalanine	0.061175	3.507418	36.71429	-0.84102	8		0.772406	0.101651		0
cyclo(leu-pro)	0.038463	3.618906	37.71429	-1.13978	8		0.094051	-1.44763		-0.42727
isoleucylisoleucine	0.14027	2.709096	15.14286	-0.91224	7		0.378602	-0.33492		0.454545
glucuronate	0.127863	2.616944	22.85714	0.288782	7		0.01426	0.60679		0.581818
X - 14255	0.09157	2.610101	26.71429	-0.66294	7		0.774773	0.114369		0.545455
4-hydroxyphenylacetyl glycine	0.09157	2.760646	27.71429	-1.04693	7		0.576191	-0.33064		0.390909
butyrylcarnitine (C4)	0.096863	3.093727	30.71429	-0.62131	7		0.649361	-0.17563		0.181818
1-oleoyl-GPE (18:1)	0.078705	2.998651	32.71429	-0.72807	7		0.724706	-0.07989		0.118182
mannose	0.048275	3.043504	34.71429	0.376952	7		0.676553	0.171845		-0.24545
hydroxyproline	0.045711	3.3184	42.71429	-0.50247	7		0.15659	-0.56452		-0.23636
3-hydroxy-2-ethylpropionate	0.038945	3.461236	46.71429	0.628257	7		0.16374	0.825209		0.5
stearoylcarnitine (C18)	0.079384	2.883895	47.71429	0.564088	7		0.028044	0.732195		0.063636
3-hydroxyoctanoate	0.027058	3.22282	53.71429	0.889034	7		0.050038	1.756323		0.8
gamma-glutamylglutamine	0.077991	3.275618	55.71429	-0.68925	7		0.232695	-0.70507		-0.18182
prolylvaline	0.079384	3.298524	61.71429	-0.66211	7		0.050038	-0.95856		-0.36364
2-hydroxybutyrate (AHB)	0.038463	2.875965	63.71429	2.274393	7		0.002873	3.04252		0.763636
nicotinamide	0.075434	2.88099	77.71429	0.63424	7		0.043385	1.346819		0.718182
oleoylcarnitine (C18)	0.038463	3.078122	84.71429	1.298099	7		0.001622	1.526236		0.290909
3-hydroxysebacate	0.038463	2.890542	87.71429	1.367667	7		0.112933	2.524536		0.454545
1-linoleoyl-GPE (18:2)	0.094466	2.545834	88.71429	-0.62334	7		0.11736	-0.92943		-0.50909
3-hydroxylaurate	0.039762	2.792687	94.71429	1.029726	7		0.001606	1.290928		0.336364
2,3-dihydroxyisovalerate	0.197227	2.658462	17.71429	0.668946	6		0.4568	0.289062		-0.54545
glutamine	0.101511	3.071818	48.71429	-0.38079	6		0.163124	-0.61918		-0.61818
N6-succinyladenosine	0.101511	3.062221	59.71429	0.409578	6		0.206133	0.512798		-0.30909
1-methylimidazoleacetate	0.125229	2.98161	83.71429	-1.11029	6		0.303727	-0.8487		0.054545
hydroxybutyrylcarnitine	0.039762	2.758099	105.7143	1.733385	6		0.057684	2.492412		0.127273
guanidinoacetate	0.07128	2.660704	121.7143	-1.09295	6		0.334217	-0.78241		-0.15455

ectoine	0.094466	2.749742	160.7143	-0.55147	6		0.218078	-0.89169		-0.28182
propionylcarnitine (C3)	0.178682	2.2971	22	-0.39652	5		0.269045	1.130749		0.890909
3-phenylpropionate (hydrocinnamate)	0.100991	2.4278	50.71429	-1.28674	5		0.502901	-0.52809		0.563636
2-hydroxyoctanoate	0.138108	2.35196	51.71429	-0.49828	5		0.13642	-0.3374		0.136364
3-hydroxy-3-methylglutarate	0.100754	2.471638	67.71429	0.811809	5		0.180597	1.082757		-0.2
N-acetyl-aspartyl-glutamate (NAAG)	0.132539	2.462324	80.71429	-0.46595	5		0.112933	-0.91901		-0.69091
X - 14473	0.129332	2.874833	102.7143	-1.20539	5		0.266483	-0.60606		-0.13636
S-methylcysteine	0.132539	2.508794	113.7143	0.397973	5		0.15659	0.508123		0.218182
myristoylcarnitine	0.078705	2.440547	154.7143	1.12988	5		0.081242	1.777995		0.318182
3-hydroxybutyrate (BHBA)	0.061585	2.339849	165.7143	1.230862	5		0.061991	1.394567		0.472727
X - 12730	0.136119	2.838823	182.7143	-1.01004	5		0.391228	-0.99791		0.645455
X - 12258	0.096635	3.220515	242.7143	-3.36777	5		0.810836	-0.15237		0.909091
2-aminoheptanoate	0.297553	2.011211	28.71429	0.455332	4		0.787946	-0.14042		-0.80909
N-(2-furoyl)glycine	0.257692	2.041109	35.71429	1.03268	4		0.75228	0.953227		-0.31818
4-imidazoleacetate	0.185748	2.472705	52.71429	-0.55443	4		0.613696	-0.15567		0.090909
indolelactate	0.155651	2.392415	54.71429	-0.56239	4		0.857033	-0.0517		0.590909
erythritol	0.184191	2.293884	68.71429	-0.28878	4		0.576191	-0.11864		0.236364
hydroquinone sulfate	0.225735	2.120301	72.71429	-0.78235	4		0.275586	0.851578		0.718182
urea	0.248103	2.146678	85.71429	0.298033	4		0.160031	0.35482		0.309091
biotin	0.185748	2.100397	95.71429	0.31421	4		0.190782	0.466264		0
glycylvaline	0.156157	2.333343	96.71429	-0.49146	4		0.500599	0.321578		0.481818
alpha-hydroxyisocaproate	0.118973	2.265921	106.7143	-0.76901	4		0.889637	0.028596		0.281818
X - 14625	0.168659	3.411513	107.7143	-0.6101	4		0.355872	-0.22648		0.127273
ferulic acid 4-sulfate	0.178682	2.680966	111.7143	-1.02735	4		0.325021	-0.94144		-0.21818
N-methyl proline	0.137196	2.410764	124.7143	-0.47963	4		0.352533	-0.25599		0.272727
adrenate (22:4n6)	0.128673	2.463403	129.7143	0.598243	4		0.38749	0.423854		0.154545
N-acetylisoleucine	0.178682	2.595007	138.7143	1.327927	4		0.047411	2.378513		0.772727

3-methyl-2-oxobutyrate	0.149683	2.188749	157.7143	0.657199	4		0.061991	0.809811		-0.00909
stearoyl-arachidonoyl-glycerophosphocholine	0.136119	2.368372	166.7143	0.21931	4		0.381855	0.238092		-0.05455
octadecanedioate (C18)	0.118973	2.313274	195.7143	1.274691	4		0.138318	2.513236		0.436364
alpha-ketobutyrate	0.094466	2.167743	201.7143	2.105077	4		0.036779	1.926085		0.081818
myristoleoylcarnitine	0.071265	2.083464	222.7143	1.580603	4		0.160031	2.889215		0.372727
16-hydroxypalmitate	0.09157	2.043567	229.7143	1.029101	4		0.041954	0.792949		-0.04545
X - 12119	0.413186	2.466847	60.71429	0.768529	3		0.345229	0.690902		0.009091
glycerol	0.172584	2.04652	108.7143	0.528179	3		0.799369	0.07492		-0.69091
N-palmitoyltaurine	0.178682	2.127812	112.7143	0.847167	3		0.02366	1.608063		0.6
N6-acetyllysine	0.168659	2.172813	114.7143	-0.2058	3		0.929051	-0.00098		0.036364
N-acetyl-1-methylhistidine	0.208394	2.058227	140.7143	0.520407	3		0.177788	0.529953		-0.03636
serine	0.18932	2.07686	141.7143	-0.27113	3		0.757872	0.245983		0.1
sphingomyelin	0.199861	2.301601	147.7143	0.377732	3		0.724706	0.098662		-0.78626
laurylcarnitine (C12)	0.122089	2.258723	234.7143	0.909095	3		0.11064	2.165485		0.6
indolebutyrate	0.132539	2.023233	244.7143	1.544192	3		0.161514	1.130501		-0.18182
X - 12199	0.219754	2.724844	267.7143	0.343264	3		0.481583	0.276295		0.418182
X - 18886	0.474621	3.574707	274.7143	-0.32743	3		0.012332	1.807576		0.509091
X - 17761	0.666353	3.830359	283.7143	-0.55439	3		0.106012	1.227942		0.109091
X - 16975	0.62471	4.373364	588.7143	0.436069	3		0.036779	-2.66541		-0.20957
X - 12721	0.503013	3.621673	690.7143	0.396174	3		0.213615	3.623974		-0.01818
spermidine	0.37129	2.016487	125.7143	-0.98018	2		0.195701	-1.55858		-0.73636
X - 14352	0.491258	2.087236	190.7143	-0.28112	2		0.446721	-0.60666		-0.15455
palmitate (16:0)	0.168659	2.073408	218.7143	0.443168	2		0.322934	0.194929		-0.29091
2-aminobutyrate	0.156157	2.032901	240.7143	1.070913	2		0.11736	1.254029		0
X - 21892	0.503013	3.319036	282.7143	0.383963	2		0.178308	1.329189		0.636364
X - 17717	0.824218	3.086305	290.7143	-0.23847	2		0.106012	-1.52571		-0.36364
X - 13835	0.776447	2.80843	359.7143	0.106016	2		0.814244	0.163213		0.181818
X - 21431	0.328006	3.037525	377.7143	0.442945	2		0.946454	0.011107		0.618182

X - 12007	0.682035	3.473447	440.7143	0.659221	2		0.14264	-2.06964		0.127563
X - 12860	0.928871	2.800688	497.7143	-0.1029	2		0.036779	1.931848		0.536364
xylitol	0.987065	2.6614	522.7143	-0.07042	2		0.943652	0.023323		0.790909
X - 14291	0.980162	2.95142	552.7143	0.296201	2		0.194355	-0.22227		0.063636
X - 14318	0.424411	3.140063	566.7143	0.71463	2		0.245896	-0.50499		0.190909
X - 17735	0.94223	2.667536	590.7143	-0.05544	2		0.094051	0.77872		0.390909
X - 18913	0.628189	2.691591	596.7143	-0.16883	2		0.559021	0.385061		0.663636
X - 12329	0.731265	3.425448	600.7143	0.162866	2		0.827012	0.174065		0.536364
X - 12267	0.92197	3.353273	644.7143	0.048881	2		0.032394	-3.0761		-0.44039
X - 21803	0.762902	3.028953	691.7143	0.152671	2		0.707045	0.716417		0.890909
X - 21295	0.777172	2.506966	731.7143	-0.15771	2		0.902744	0.077049		0.518182
X - 21353	0.987065	2.578551	740.7143	-0.03502	2		0.008429	1.385294		0.409091
X - 11441	0.97301	2.562544	778.7143	-0.01306	2		0.11736	-0.97172		-0.51818
X - 21792	0.375741	2.384679	215.7143	-0.06768	1		0.135949	0.769303		-0.57273
X - 21729	0.492163	2.482178	231.7143	-0.31958	1		0.994976	0.100088		0.5
X - 14374	0.814386	2.093024	345.7143	-0.10476	1		0.269045	-0.61909		-0.43636
X - 12095	0.905322	2.316029	354.7143	-0.04393	1		0.178308	1.407377		0.809091
X - 12212	0.628189	2.307694	427.7143	0.220189	1		0.951991	0.00571		0.590909
X - 13743	0.979982	2.184976	507.7143	0.061067	1		0.061991	2.316824		0.581818

Table 3. Metabolites identified as significantly different in the liver of the low dose rats after 4 hours.

	Low dose vs Control						High dose vs Control			Exposure correlation
Metabolites	**q-value**	**VIP Score**	**RFECV mean**	**Fold-change (log2)**	**Cons. level**		**q-value**	**Fold-change (log2)**		**Rho**
AMP	0.251345	4.647859	1	0.575549	7		0.155969	0.456934		0.036364
3-hydroxy-3-methylglutarate	0.251345	4.537536	1	1.414274	7		0.131955	1.071386		-0.4
adenylosuccinate	0.251345	4.040677	1	1.839679	7		0.251481	1.836392		0.255126
serine	0.321991	5.909512	1	0.353674	6		0.015601	0.681229		0.754545
pyruvate	0.469938	4.766416	1	-1.3947	6		0.353482	-2.27927		-0.26364
erythritol	0.484886	4.476072	1	-0.51379	6		0.258664	-0.56683		0.01373
maltopentaose	0.484886	4.363443	24.57143	-1.29013	6		0.345109	-1.39642		-0.09567
maltotriose	0.518649	4.352898	16.42857	-1.06904	6		0.109164	-4.97111		-0.7
gamma-glutamyltryptophan	0.399456	4.206596	1	0.92859	6		0.499991	0.459053		-0.48182
stearoyl-arachidonoyl-glycerophosphoinositol	0.484886	4.203953	1	-0.21669	6		0.536955	-0.16925		0.172727
2-aminophenol sulfate	0.484886	4.102025	2.285714	-0.83889	6		0.962405	0.024839		0.618182
butyrylcarnitine (C4)	0.321991	4.091793	1	-0.95705	6		0.130422	-1.2094		0.136364
ribose 1-phosphate	0.484886	4.057763	1	0.327199	6		0.464253	0.256072		0.036364
guanidinosuccinate	0.469938	4.040342	6.857143	0.719335	6		0.581394	0.421889		-0.26364
NAD+	0.37723	4.018621	4.285714	0.670579	6		0.249514	0.545851		-0.00909
glycerol 3-phosphate (G3P)	0.518649	3.759778	1.285714	0.643599	6		0.248532	0.751303		0.236364
pregnenolone sulfate	0.510741	3.753985	13.57143	1.718983	6		0.045742	2.583251		0.818182
gamma-glutamylglutamate	0.484886	3.738857	1	0.556691	6		0.654423	0.49367		-0.11818
anserine	0.484886	3.738824	1	0.81785	6		0.169926	0.719984		0
tigloylglycine	0.510741	3.632585	5.142857	0.558072	6		0.054162	1.263174		0.563636
pregnanolone/allopregnanolone sulfate	0.484886	3.520798	1	0.942781	6		0.335745	1.307445		0.145455

X - 12101	0.836272	4.276358	58.57143	0.593844	5		0.755883	0.532448		-0.3
sorbitol	0.484886	3.877457	90.57143	-0.69679	5		0.102996	-1.99862		-0.63636
formononetin	0.470893	3.73319	37.57143	-1.05423	5		0.783819	-0.0889		0.3
S-methylcysteine	0.470893	3.577633	32.57143	0.613019	5		0.027878	0.757693		0.309091
pantethine	0.321991	3.549615	49.57143	1.090461	5		0.166789	1.690908		0.336364
choline	0.459952	3.527017	26.57143	0.593089	5		0.103076	1.116603		0.672727
deoxycarnitine	0.484886	3.467673	1.142857	-0.36654	5		0.656033	0.175132		0.590909
phenyllactate (PLA)	0.484886	3.394419	22.57143	-1.77237	5		0.515439	-1.15963		0.396356
3-hydroxy-2-ethylpropionate	0.518649	3.341656	10	0.495403	5		0.142458	0.76622		0.536364
N-acetyl-glucosamine 1-phosphate	0.484886	3.304682	11.42857	0.313726	5		0.027878	0.366046		0.118182
1-docosapentaenoyl-GPC (22:5n6)	0.484886	3.24286	3.428571	-1.09815	5		0.602806	1.299314		0.564922
2'-O-methylguanosine	0.484886	3.237808	8.142857	0.556776	5		0.261862	1.165277		0.363636
2-linoleoylglycerophosphoinositol*	0.57675	3.154069	3.857143	0.789474	5		0.984572	0.192074		-0.3
ribonate (ribonolactone)	0.585049	3.128943	3	-0.37896	5		0.27445	-0.73493		-0.48182
alpha-hydroxyisovaleroyl carnitine	0.510741	3.094648	2	0.266759	5		0.015601	0.813819		0.918182
orotate	0.510741	3.070479	5.571429	0.450994	5		0.454171	0.373604		-0.08182
methyl glucopyranoside (alpha + beta)	0.484886	3.030993	6	-0.56832	5		0.126194	-1.37705		-0.69091
ethylmalonate	0.487825	2.909813	1.428571	-0.84504	5		0.158057	-1.32079		-0.14545
glycerophosphoethanolamine	0.484886	2.786142	10.71429	-0.41272	5		0.093721	-1.15146		-0.54545
gamma-glutamyltyrosine	0.510741	2.57867	6.428571	0.958499	5		0.253946	1.272921		-0.1
X - 21861	0.687161	3.960112	126.5714	0.47938	4		0.015601	1.018282		0.481818
pseudouridine	0.484886	3.345503	27.57143	0.318074	4		0.187311	0.456538		0.372727
fructose	0.484886	3.282759	68.57143	-0.51918	4		0.075431	-1.67453		-0.53636
pipecolate	0.484886	3.251177	55.57143	0.525826	4		0.296993	0.530051		0.009091

sedoheptulose-7-phosphate	0.321991	3.130161	39.57143	1.272271	4		0.109164	1.98693		0.645455
2-hydroxypalmitate	0.484886	3.126327	36.57143	0.340515	4		0.877241	-0.0672		-0.52727
cyclo(leu-pro)	0.598389	3.093743	43.57143	-0.63657	4		0.126194	-1.47636		-0.69091
daidzein	0.484886	3.093106	67.57143	-0.94926	4		0.976798	0.279671		0.572727
phosphoenolpyruvate (PEP)	0.484886	3.011915	77.57143	0.558739	4		0.060034	0.795191		0.327273
UMP	0.484886	3.004034	84.57143	0.718771	4		0.499991	0.505116		-0.02727
3-hydroxybutyrate (BHBA)	0.484886	2.986459	75.57143	0.546377	4		0.039742	1.020909		0.645455
beta-alanine	0.484886	2.982172	70.57143	0.389176	4		0.314276	0.913213		0.463636
2-aminobutyrate	0.485138	2.972555	86.57143	1.074673	4		0.126194	1.261711		-0.09091
N-acetylglucosamine 6-phosphate	0.484886	2.938438	34.57143	0.2677	4		0.057727	0.379176		0.354545
X - 14838	0.734151	2.914205	31.57143	-0.22631	4		0.247931	-1.18429		-0.09091
N-acetylserine	0.518649	2.888999	72.57143	0.394878	4		0.166789	0.594446		0.309091
caprylate (8:0)	0.592595	2.807356	28.57143	0.249104	4		0.594081	0.114608		-0.30909
isobutyrylglycine (C4)	0.592595	2.761467	79.57143	0.87706	4		0.651308	0.921503		0.095672
palmitoyl-linoleoyl-glycerophosphocholine	0.531951	2.717766	50.57143	0.266403	4		0.503613	0.174506		-0.53636
glutaroylcarnitine (C5)	0.597432	2.713239	76.57143	-0.98269	4		0.258664	-1.1943		-0.07273
gamma-glutamylvaline	0.484886	2.686069	57.57143	-0.97001	4		0.078486	-1.66078		-0.49887
uridine-3'-monophosphate (3'-UMP)	0.592595	2.571951	30.57143	-0.90861	4		0.258664	-1.09788		0.050114
glycylmethionine	0.518649	2.558059	52.57143	0.295096	4		0.875022	-0.0801		-0.01818
7-methylguanine	0.592595	2.543879	83.57143	0.194896	4		0.314276	0.484358		0.545455
naringenin	0.592595	2.457126	18.14286	2.167313	4		0.251481	0.75517		-0.27273
5-hydroxyindoleacetate	0.594136	2.371931	19.85714	-0.38192	4		0.83664	-0.12133		0.509091
2-palmitoleoylglycerophosphoinositol	0.615349	2.367359	1.714286	-0.43467	4		0.474451	-0.36813		0.214124
stachydrine	0.798062	2.351723	4.714286	0.214314	4		0.65782	-0.17717		-0.54545
2-palmitoyl-GPE (16:0)	0.728062	2.315985	17.28571	-0.40951	4		0.617505	1.028252		0.436364

gamma-glutamylphenylalanine	0.578435	2.238667	15.71429	0.519822	4		0.598742	0.410353		-0.2
beta-muricholate	0.839197	2.085134	9.285714	-1.87281	4		0.281735	1.309749		0.581818
X - 12104	0.960735	5.904875	560.5714	0.067944	3		0.027878	1.515314		0.381818
X - 21343	0.900806	4.931907	511.5714	-0.16382	3		0.251481	0.245576		-0.43636
X - 13737	0.839722	4.294123	235.5714	-0.2114	3		0.256536	1.255444		0.390909
X - 12442	0.689867	4.242108	350.5714	0.552862	3		0.166789	0.654625		0.2
X - 11979	0.992882	3.764186	663.5714	-0.48265	3		0.126194	1.344173		0.6
xanthosine	0.959309	3.676239	595.5714	0.084041	3		0.126194	-0.86779		-0.28182
maltose	0.518649	3.39027	138.5714	-0.6364	3		0.137782	-3.07137		-0.65455
threonine	0.484886	3.125971	111.5714	0.360546	3		0.014594	0.962483		0.845455
ferulic acid 4-sulfate	0.510741	2.880105	129.5714	-1.07817	3		0.503613	-1.38788		-0.27273
cytidine	0.484886	2.84422	161.5714	0.162603	3		0.169926	0.506487		0.572727
X - 18938	0.801777	2.681402	163.5714	0.370132	3		0.985417	-0.00639		0.472727
N6-carbamoylthreonyladenosine	0.518649	2.660934	118.5714	0.37195	3		0.06775	0.677222		0.772727
X - 17246	0.942126	2.652088	143.5714	0.278139	3		0.503613	-0.68233		0.405468
CMP	0.484886	2.604571	174.5714	1.069898	3		0.007879	1.88241		0.590909
X - 21755	0.743965	2.599205	181.5714	-1.06348	3		0.023252	0.408708		0.309091
N-acetylarginine	0.510741	2.554252	136.5714	0.412836	3		0.055289	1.008054		0.818182
benzoate	0.54495	2.55264	109.5714	0.205229	3		0.949944	-0.02686		-0.60909
glutamine	0.510741	2.547358	104.5714	-0.26093	3		0.039742	-0.72482		-0.79091
dihydrokaempferol	0.532138	2.498276	38.57143	3.233653	3		0.186444	1.07412		-0.24545
5-hydroxytryptophol	0.518649	2.491922	29.57143	-0.49425	3		0.406778	-0.42535		0.2
glycyltyrosine	0.518649	2.371132	56.57143	0.305277	3		0.822352	-0.09339		-0.01818
phosphate	0.531951	2.359067	74.57143	0.088451	3		0.919087	0.045273		0.327273
N-acetylaspartate (NAA)	0.57675	2.285455	45.57143	-0.84202	3		0.207473	-1.05378		-0.32727
glycerophosphorylcholine (GPC)	0.518649	2.27612	80.57143	-0.56697	3		0.057727	-1.37749		-0.57273

UDP-galactose	0.551652	2.170553	96.57143	0.569762	3		0.609184	-0.40655		-0.58182
2-hydroxyhippurate (salicylurate)	0.828417	2.167493	65.57143	1.284875	3		0.533898	1.024062		-0.24715
4-imidazoleacetate	0.757001	2.101049	48.57143	-0.39225	3		0.362345	-0.69866		-0.13636
X - 22768	0.484886	2.035546	99.57143	-1.31353	3		0.54833	1.62861		0.290909
xanthine	0.782978	2.00367	87.57143	0.348489	3		0.354236	0.207761		0.427273
X - 21353	0.942126	3.401222	454.5714	-0.16207	2		0.845096	0.109469		-0.48182
xylitol	0.998209	3.248118	376.5714	-0.31138	2		0.251172	-1.30637		-0.52727
X - 16944	0.896415	3.179174	783.5714	0.196922	2		0.327855	0.517264		0
X - 12125	0.998209	3.172699	265.5714	0.037449	2		0.093721	0.83474		0.709091
X - 12695	0.959309	3.027875	795.5714	0.305551	2		0.09595	0.994052		0.818182
X - 17694	0.998209	3.000108	624.5714	-0.09377	2		0.542518	0.744688		-0.15597
X - 11576	0.900806	2.912117	390.5714	-0.03535	2		0.241714	0.545387		0.018182
X - 12689	0.900806	2.897862	707.5714	0.133537	2		0.521991	-0.23952		0.209091
X - 14961	0.959607	2.859556	453.5714	-0.05858	2		0.231762	-1.36676		-0.60909
X - 11360	0.782978	2.777293	366.5714	-0.30814	2		0.280395	-1.25456		-0.35455
trigonelline (N'-methylnicotinate)	0.510741	2.528682	204.5714	-0.35254	2		0.253946	-1.12385		-0.55455
CDP-ethanolamine	0.510741	2.416931	130.5714	0.63547	2		0.674133	0.469107		0.118182
alpha-ketoglutarate	0.510741	2.348807	188.5714	-0.61884	2		0.093721	-1.03308		-0.07273
5-dodecenoate (12:1n7)	0.54495	2.340059	128.5714	0.546371	2		0.258664	1.35749		0.436364
guanosine	0.54495	2.327632	122.5714	0.208094	2		0.899078	-0.02414		-0.02727
oleic ethanolamide	0.55335	2.110588	180.5714	0.354897	2		0.542518	0.325099		-0.07273
myristoleoylcarnitine	0.592595	2.100654	168.5714	-0.63677	2		0.109164	-1.21696		-0.43636
glutathione, reduced (GSH)	0.592595	2.016793	125.5714	-0.2155	2		0.015601	-1.1759		-0.66364
mannose	0.592595	2.395832	242.5714	-0.41458	1		0.057727	-1.79364		-0.55455
X - 12860	0.896415	2.385132	330.5714	0.191399	1		0.241389	-1.30728		-0.05455
androsterone sulfate	0.599334	2.342602	316.5714	0.685294	1		0.027878	1.617424		0.827273
X - 22162	0.950879	2.306493	219.5714	-0.46876	1		0.917436	0.081696		-0.21818

X - 17709	0.998209	2.306	317.5714	-0.71005	1		0.057727	1.224718		0.536364
X - 22585	0.7363	2.303402	809.5714	-0.30763	1		0.231762	-1.28963		0.281133
hexadecanedioate (C16)	0.484886	2.258314	249.5714	0.586457	1		0.101606	1.155971		0.827273
valine	0.531951	2.194454	202.5714	0.193278	1		0.131955	0.334574		0.627273
X - 21444	0.896415	2.176711	208.5714	0.274327	1		0.123197	-2.48539		-0.62244
X - 21607	0.518649	2.158387	395.5714	-0.863	1		0.542518	0.304174		-0.43636
alanine	0.549772	2.145709	215.5714	-0.13299	1		0.549773	0.316233		0.6
glucose	0.597432	2.128497	261.5714	-0.30598	1		0.126194	-1.64074		-0.6
X - 14626	0.926492	2.06332	548.5714	0.109353	1		0.919087	0.126464		0.190909

Appendix II: The novel predicted mitochondrial proteins

Gene ID	Symbol	Probability		Gene ID	Symbol	Probability
ENSG00000177150	FAM210A	1		ENSG00000165443	PHYHIPL	0.921197
ENSG00000184857	TMEM186	1		ENSG00000112208	BAG2	0.920945
ENSG00000137274	BPHL	1		ENSG00000205002	AARD	0.920675
ENSG00000157326	DHRS4	1		ENSG00000126947	ARMCX1	0.920409
ENSG00000050426	LETMD1	1		ENSG00000206052	DOK6	0.919912
ENSG00000172992	DCAKD	1		ENSG00000100422	CERK	0.919119
ENSG00000130748	TMEM160	1		ENSG00000089597	GANAB	0.91765
ENSG00000142444	C19orf52	1		ENSG00000257727	CNPY2	0.916305
ENSG00000159348	CYB5R1	1		ENSG00000166199	ALKBH3	0.915694
ENSG00000160439	RDH13	1		ENSG00000072133	RPS6KA6	0.914952
ENSG00000130349	C6orf203	1		ENSG00000148730	EIF4EBP2	0.914561
ENSG00000168393	DTYMK	1		ENSG00000131379	C3orf20	0.91412
ENSG00000161558	TMEM143	1		ENSG00000105135	ILVBL	0.911377
ENSG00000178096	BOLA1	1		ENSG00000084733	RAB10	0.910815
ENSG00000173137	ADCK5	1		ENSG00000134824	FADS2	0.910214
ENSG00000180011	ZADH2	1		ENSG00000174444	RPL4	0.910017
ENSG00000116096	SPR	1		ENSG00000184840	TMED9	0.909878
ENSG00000156398	SFXN2	1		ENSG00000147403	RPL10	0.909084
ENSG00000184227	ACOT1	1		ENSG00000186591	UBE2H	0.908941
ENSG00000114021	NIT2	1		ENSG00000111647	UHRF1BP1L	0.90814
ENSG00000117528	ABCD3	1		ENSG00000159374	M1AP	0.908099
ENSG00000164241	C5orf63	1		ENSG00000189283	FHIT	0.906965
ENSG00000165792	METTL17	1		ENSG00000114942	EEF1B2	0.906159
ENSG00000063761	ADCK1	1		ENSG00000105193	RPS16	0.906126
ENSG00000165028	NIPSNAP3B	1		ENSG00000135046	ANXA1	0.906087

ENSG00000205544	TMEM256	1		ENSG00000255154	RPP14	0.905946
ENSG00000164040	PGRMC2	1		ENSG00000124614	RPS10	0.905404
ENSG00000100714	MTHFD1	1		ENSG00000114573	ATP6V1A	0.904648
ENSG00000114735	HEMK1	0.999999		ENSG00000124207	CSE1L	0.903871
ENSG00000204394	VARS	0.999999		ENSG00000250317	SMIM20	0.903199
ENSG00000100445	SDR39U1	0.999999		ENSG00000105640	RPL18A	0.90169
ENSG00000120992	LYPLA1	0.999999		ENSG00000204628	GNB2L1	0.900423
ENSG00000122378	FAM213A	0.999999		ENSG00000152904	GGPS1	0.899981
ENSG00000008394	MGST1	0.999998		ENSG00000134285	FKBP11	0.89986
ENSG00000164924	YWHAZ	0.999997		ENSG00000162496	DHRS3	0.89904
ENSG00000163866	SMIM12	0.999997		ENSG00000165152	TMEM246	0.898951
ENSG00000163607	GTPBP8	0.999997		ENSG00000140798	ABCC12	0.898718
ENSG00000112304	ACOT13	0.999996		ENSG00000156482	RPL30	0.896773
ENSG00000147592	LACTB2	0.999996		ENSG00000179869	ABCA13	0.896169
ENSG00000231500	RPS18	0.999996		ENSG00000133121	STARD13	0.895773
ENSG00000133597	ADCK2	0.999995		ENSG00000163617	KIAA1407	0.895589
ENSG00000180488	FAM73A	0.999995		ENSG00000068784	SRBD1	0.895386
ENSG00000167004	PDIA3	0.999994		ENSG00000244694	PTCHD4	0.894809
ENSG00000165233	C9orf89	0.999994		ENSG00000138760	SCARB2	0.894471
ENSG00000167264	DUS2	0.999992		ENSG00000166133	RPUSD2	0.893829
ENSG00000197746	PSAP	0.99999		ENSG00000183891	TTC32	0.893413
ENSG00000157379	DHRS1	0.99999		ENSG00000075035	WSCD2	0.893262
ENSG00000139990	DCAF5	0.999989		ENSG00000183569	SERHL2	0.892904
ENSG00000117450	PRDX1	0.999987		ENSG00000124194	GDAP1L1	0.891744
ENSG00000186603	HPDL	0.999984		ENSG00000203859	HSD3B2	0.891547
ENSG00000067225	PKM	0.999983		ENSG00000006451	RALA	0.890808
ENSG00000204564	C6orf136	0.999983		ENSG00000127526	SLC35E1	0.890049

ENSG00000170634	ACYP2	0.997425		ENSG00000147604	RPL7	0.889564
ENSG00000060971	ACAA1	0.997353		ENSG00000113643	RARS	0.889308
ENSG00000204237	OXLD1	0.996793		ENSG00000100030	MAPK1	0.887955
ENSG00000150787	PTS	0.99677		ENSG00000068097	HEATR6	0.887185
ENSG00000254402	LRRC24	0.996748		ENSG00000141338	ABCA8	0.887132
ENSG00000110011	DNAJC4	0.996745		ENSG00000174903	RAB1B	0.887105
ENSG00000170889	RPS9	0.996552		ENSG00000186314	PRELID2	0.887088
ENSG00000089157	RPLP0	0.996277		ENSG00000122406	RPL5	0.886808
ENSG00000168273	SMIM4	0.995782		ENSG00000129151	BBOX1	0.886556
ENSG00000166598	HSP90B1	0.995767		ENSG00000244038	DDOST	0.886205
ENSG00000205707	LYRM5	0.995506		ENSG00000197959	DNM3	0.885691
ENSG00000109971	HSPA8	0.995223		ENSG00000039560	RAI14	0.884775
ENSG00000132570	PCBD2	0.995045		ENSG00000025708	TYMP	0.88475
ENSG00000198755	RPL10A	0.994871		ENSG00000099977	DDT	0.884512
ENSG00000160752	FDPS	0.994764		ENSG00000155890	TRIM42	0.88284
ENSG00000172270	BSG	0.994757		ENSG00000114383	TUSC2	0.882763
ENSG00000073169	SELO	0.994729		ENSG00000117016	RIMS3	0.882585
ENSG00000026025	VIM	0.994312		ENSG00000063177	RPL18	0.882197
ENSG00000142168	SOD1	0.993291		ENSG00000142459	EVI5L	0.882088
ENSG00000168569	TMEM223	0.992672		ENSG00000211456	SACM1L	0.881837
ENSG00000101166	SLMO2	0.992441		ENSG00000071082	RPL31	0.881441
ENSG00000161016	RPL8	0.992206		ENSG00000215301	DDX3X	0.879846
ENSG00000184524	CEND1	0.992131		ENSG00000077157	PPP1R12B	0.879218
ENSG00000224877	C17orf89	0.992103		ENSG00000164645	C7orf62	0.878521
ENSG00000170791	CHCHD7	0.992001		ENSG00000093010	COMT	0.876406
ENSG00000084207	GSTP1	0.991264		ENSG00000130707	ASS1	0.876376
ENSG00000110719	TCIRG1	0.991186		ENSG00000092841	MYL6	0.876238

ENSG00000154258	ABCA9	0.99081		ENSG00000161203	AP2M1	0.876179
ENSG00000100316	RPL3	0.990524		ENSG00000166833	NAV2	0.875443
ENSG00000159596	TMEM69	0.990339		ENSG00000239672	NME1	0.874996
ENSG00000186081	KRT5	0.990305		ENSG00000080824	HSP90AA1	0.874615
ENSG00000179988	PSTK	0.99006		ENSG00000174177	CTU2	0.87374
ENSG00000006125	AP2B1	0.989959		ENSG00000067560	RHOA	0.872881
ENSG00000160883	HK3	0.989361		ENSG00000140284	SLC27A2	0.872674
ENSG00000204427	ABHD16A	0.989176		ENSG00000204653	ASPDH	0.871421
ENSG00000198650	TAT	0.989152		ENSG00000167641	PPP1R14A	0.871216
ENSG00000184992	BRI3BP	0.988635		ENSG00000128245	YWHAH	0.870473
ENSG00000118363	SPCS2	0.988624		ENSG00000104059	FAM189A1	0.870027
ENSG00000011052	NME2	0.988411		ENSG00000186298	PPP1CC	0.869349
ENSG00000110917	MLEC	0.988314		ENSG00000277791	PSMB3	0.86901
ENSG00000164398	ACSL6	0.988295		ENSG00000170921	TANC2	0.868841
ENSG00000102794	IRG1	0.988192		ENSG00000116337	AMPD2	0.868054
ENSG00000165553	NGB	0.98798		ENSG00000188343	FAM92A1	0.867584
ENSG00000149428	HYOU1	0.987872		ENSG00000186051	TAL2	0.86729
ENSG00000111640	GAPDH	0.98779		ENSG00000112514	CUTA	0.866533
ENSG00000136628	EPRS	0.98777		ENSG00000100364	KIAA0930	0.866479
ENSG00000135362	PRR5L	0.987471		ENSG00000065060	UHRF1BP1	0.865823
ENSG00000127948	POR	0.986887		ENSG00000127452	FBXL12	0.86544
ENSG00000109475	RPL34	0.986424		ENSG00000198610	AKR1C4	0.86419
ENSG00000197157	SND1	0.985957		ENSG00000146350	TBC1D32	0.863969
ENSG00000196230	TUBB	0.985616		ENSG00000138101	DTNB	0.861792
ENSG00000047230	CTPS2	0.985431		ENSG00000085662	AKR1B1	0.861765
ENSG00000132763	MMACHC	0.985112		ENSG00000188483	IER5L	0.861742
ENSG00000096384	HSP90AB1	0.984982		ENSG00000130734	ATG4D	0.861304

ENSG00000100596	SPTLC2	0.98497		ENSG00000144369	FAM171B	0.860189
ENSG00000162398	C1orf177	0.984928		ENSG00000197045	GMFB	0.859941
ENSG00000167815	PRDX2	0.984792		ENSG00000075785	RAB7A	0.85919
ENSG00000119917	IFIT3	0.983932		ENSG00000197226	TBC1D9B	0.858618
ENSG00000119943	PYROXD2	0.983685		ENSG00000131370	SH3BP5	0.857403
ENSG00000172482	AGXT	0.983442		ENSG00000197006	METTL9	0.857143
ENSG00000074800	ENO1	0.983232		ENSG00000080371	RAB21	0.856143
ENSG00000185386	MAPK11	0.982913		ENSG00000122203	KIAA1191	0.85566
ENSG00000108828	VAT1	0.982846		ENSG00000118520	ARG1	0.854239
ENSG00000182899	RPL35A	0.982816		ENSG00000117592	PRDX6	0.853315
ENSG00000164978	NUDT2	0.982691		ENSG00000244187	TMEM141	0.853132
ENSG00000164587	RPS14	0.981636		ENSG00000109046	WSB1	0.853024
ENSG00000134419	RPS15A	0.980925		ENSG00000182718	ANXA2	0.852473
ENSG00000240857	RDH14	0.980504		ENSG00000162927	PUS10	0.851334
ENSG00000137038	TMEM261	0.979148		ENSG00000106772	PRUNE2	0.851187
ENSG00000181915	ADO	0.978351		ENSG00000138801	PAPSS1	0.84898
ENSG00000103502	CDIPT	0.978048		ENSG00000135821	GLUL	0.848464
ENSG00000166348	USP54	0.977858		ENSG00000124786	SLC35B3	0.847544
ENSG00000166347	CYB5A	0.977237		ENSG00000173540	GMPPB	0.846092
ENSG00000111669	TPI1	0.975584		ENSG00000130589	HELZ2	0.846053
ENSG00000167654	ATCAY	0.97504		ENSG00000104412	EMC2	0.845244
ENSG00000148343	FAM73B	0.974711		ENSG00000125691	RPL23	0.845114
ENSG00000197043	ANXA6	0.974328		ENSG00000105784	RUNDC3B	0.843567
ENSG00000143554	SLC27A3	0.973068		ENSG00000140280	LYSMD2	0.842684
ENSG00000103254	FAM173A	0.973064		ENSG00000088320	REM1	0.842581
ENSG00000164163	ABCE1	0.972959		ENSG00000139438	FAM222A	0.842521
ENSG00000166340	TPP1	0.972858		ENSG00000086289	EPDR1	0.842086

ENSG00000041988	THAP3	0.972778		ENSG00000167996	FTH1	0.842018
ENSG00000141367	CLTC	0.972258		ENSG00000074755	ZZEF1	0.841716
ENSG00000177600	RPLP2	0.971742		ENSG00000105726	ATP13A1	0.840398
ENSG00000133687	TMTC1	0.970902		ENSG00000166329	CCDC182	0.839901
ENSG00000135002	RFK	0.970585		ENSG00000119771	KLHL29	0.83976
ENSG00000033050	ABCF2	0.969855		ENSG00000089127	OAS1	0.839752
ENSG00000138363	ATIC	0.96947		ENSG00000144713	RPL32	0.839209
ENSG00000106436	MYL10	0.966689		ENSG00000120705	ETF1	0.838718
ENSG00000101152	DNAJC5	0.966505		ENSG00000120697	ALG5	0.838158
ENSG00000232859	LYRM9	0.966441		ENSG00000184867	ARMCX2	0.838031
ENSG00000135116	HRK	0.966428		ENSG00000105372	RPS19	0.837808
ENSG00000134308	YWHAQ	0.966381		ENSG00000157999	ANKRD61	0.836447
ENSG00000087086	FTL	0.965572		ENSG00000111790	FGFR1OP2	0.83632
ENSG00000115275	MOGS	0.964986		ENSG00000196872	KIAA1211L	0.835759
ENSG00000110497	AMBRA1	0.964499		ENSG00000125844	RRBP1	0.834867
ENSG00000101337	TM9SF4	0.963812		ENSG00000162777	DENND2D	0.83473
ENSG00000122884	P4HA1	0.963555		ENSG00000165661	QSOX2	0.834581
ENSG00000067177	PHKA1	0.963552		ENSG00000167645	YIF1B	0.834465
ENSG00000171097	CCBL1	0.962574		ENSG00000107099	DOCK8	0.830902
ENSG00000072274	TFRC	0.962385		ENSG00000158792	SPATA2L	0.830852
ENSG00000101473	ACOT8	0.961463		ENSG00000135624	CCT7	0.830115
ENSG00000173681	CXorf23	0.961209		ENSG00000115464	USP34	0.829579
ENSG00000165632	TAF3	0.960227		ENSG00000214827	MTCP1	0.829559
ENSG00000074695	LMAN1	0.959058		ENSG00000196743	GM2A	0.829065
ENSG00000213593	TMX2	0.958735		ENSG00000131149	GSE1	0.82884
ENSG00000036448	MYOM2	0.958655		ENSG00000102882	MAPK3	0.828612
ENSG00000164488	DACT2	0.957867		ENSG00000119285	HEATR1	0.828596

ENSG00000177692	DNAJC28	0.957604		ENSG00000167552	TUBA1A	0.828343
ENSG00000142937	RPS8	0.957554		ENSG00000136682	CBWD2	0.827447
ENSG00000254772	EEF1G	0.957049		ENSG00000143353	LYPLAL1	0.826878
ENSG00000221886	ZBED8	0.956369		ENSG00000156510	HKDC1	0.826857
ENSG00000117480	FAAH	0.956036		ENSG00000066739	ATG2B	0.826703
ENSG00000086061	DNAJA1	0.954798		ENSG00000174021	GNG5	0.825851
ENSG00000153982	GDPD1	0.954667		ENSG00000250067	YJEFN3	0.825764
ENSG00000166794	PPIB	0.95432		ENSG00000078369	GNB1	0.825535
ENSG00000160285	LSS	0.95339		ENSG00000109846	CRYAB	0.825319
ENSG00000111906	HDDC2	0.952805		ENSG00000168765	GSTM4	0.824858
ENSG00000188738	FSIP2	0.951657		ENSG00000182774	RPS17	0.82484
ENSG00000141391	SLMO1	0.951298		ENSG00000141401	IMPA2	0.824746
ENSG00000186150	UBL4B	0.950763		ENSG00000137198	GMPR	0.824092
ENSG00000213719	CLIC1	0.95005		ENSG00000204311	DFNB59	0.82355
ENSG00000162032	SPSB3	0.949877		ENSG00000132530	XAF1	0.822391
ENSG00000140107	SLC25A47	0.949552		ENSG00000155792	DEPTOR	0.821959
ENSG00000100612	DHRS7	0.949288		ENSG00000139641	ESYT1	0.821331
ENSG00000104723	TUSC3	0.949219		ENSG00000156958	GALK2	0.821226
ENSG00000167588	GPD1	0.94754		ENSG00000007516	BAIAP3	0.820643
ENSG00000120265	PCMT1	0.947407		ENSG00000116151	MORN1	0.820643
ENSG00000151640	DPYSL4	0.947395		ENSG00000100991	TRPC4AP	0.820107
ENSG00000101986	ABCD1	0.946485		ENSG00000137968	SLC44A5	0.820083
ENSG00000072110	ACTN1	0.945643		ENSG00000099797	TECR	0.819986
ENSG00000060982	BCAT1	0.945034		ENSG00000117114	ADGRL2	0.818956
ENSG00000179933	C14orf119	0.944973		ENSG00000135632	SMYD5	0.818744
ENSG00000074696	HACD3	0.944727		ENSG00000215712	TMEM242	0.818363
ENSG00000198682	PAPSS2	0.943718		ENSG00000124164	VAPB	0.818098

ENSG00000170899	GSTA4	0.943117		ENSG00000173486	FKBP2	0.817939
ENSG00000171433	GLOD5	0.942852		ENSG00000162813	BPNT1	0.817677
ENSG00000134884	ARGLU1	0.942776		ENSG00000215845	TSTD1	0.817497
ENSG00000104763	ASAH1	0.942583		ENSG00000182676	PPP1R27	0.817361
ENSG00000118705	RPN2	0.941903		ENSG00000138413	IDH1	0.816266
ENSG00000109511	ANXA10	0.941382		ENSG00000114857	NKTR	0.8151
ENSG00000134108	ARL8B	0.938531		ENSG00000130304	SLC27A1	0.814207
ENSG00000182712	CMC4	0.938398		ENSG00000225921	NOL7	0.81392
ENSG00000163902	RPN1	0.938157		ENSG00000179314	WSCD1	0.813711
ENSG00000106346	USP42	0.93783		ENSG00000164124	TMEM144	0.813297
ENSG00000178927	C17orf62	0.93639		ENSG00000107518	ATRNL1	0.812071
ENSG00000180879	SSR4	0.936366		ENSG00000144867	SRPRB	0.811645
ENSG00000103512	NOMO1	0.935814		ENSG00000114374	USP9Y	0.8115
ENSG00000109016	DHRS7B	0.934397		ENSG00000163596	ICA1L	0.811444
ENSG00000139324	TMTC3	0.934224		ENSG00000124374	PAIP2B	0.811163
ENSG00000164327	RICTOR	0.933976		ENSG00000125170	DOK4	0.81105
ENSG00000132781	MUTYH	0.932636		ENSG00000065413	ANKRD44	0.810462
ENSG00000137700	SLC37A4	0.932539		ENSG00000154229	PRKCA	0.810284
ENSG00000139644	TMBIM6	0.932427		ENSG00000153721	CNKSR3	0.810249
ENSG00000118492	ADGB	0.931948		ENSG00000187862	TTC24	0.808903
ENSG00000114993	RTKN	0.931436		ENSG00000148225	WDR31	0.808847
ENSG00000170348	TMED10	0.930374		ENSG00000148057	IDNK	0.808832
ENSG00000137393	RNF144B	0.930291		ENSG00000130255	RPL36	0.808065
ENSG00000108953	YWHAE	0.929957		ENSG00000167526	RPL13	0.80798
ENSG00000127314	RAP1B	0.929554		ENSG00000167658	EEF2	0.807083
ENSG00000133878	DUSP26	0.92887		ENSG00000154124	OTULIN	0.807051
ENSG00000146731	CCT6A	0.928632		ENSG00000116685	KIAA2013	0.804838

ENSG00000176894	PXMP2	0.928023		ENSG00000162755	KLHDC9	0.804455
ENSG00000177889	UBE2N	0.925532		ENSG00000110700	RPS13	0.803322
ENSG00000089009	RPL6	0.925493		ENSG00000168904	LRRC28	0.803098
ENSG00000180891	CUEDC1	0.924334		ENSG00000109270	LAMTOR3	0.802992
ENSG00000142544	CTU1	0.924172		ENSG00000136854	STXBP1	0.801646
ENSG00000088854	C20orf194	0.923799		ENSG00000109436	TBC1D9	0.801477
ENSG00000154237	LRRK1	0.922141		ENSG00000128595	CALU	0.801396

Appendix III: The novel predicted mitochondrial disease genes

UniProt	Symbol	Binary disease	Multiclass disease sum	Multiclass recessive	Multiclass dominant	Multiclass other
P36873	PPP1CC	1	0.999999931	1.16E-11	0.002197631	0.9978023
P40763	STAT3	1	0.99999996	1.11E-14	0.00025596	0.999744
Q02878	RPL6	1	0.98887971	0.01626507	0.31585544	0.6567592
P12931	SRC	1	0.999999953	7.12E-20	3.65E-06	0.9999963
P26439	HSD3B2	1	0.999999888	2.19E-08	0.007930366	0.9920695
P15121	AKR1B1	1	0.999999958	1.16E-09	0.002607157	0.9973928
P36578	RPL4	1	0.998887185	0.04452034	0.121593095	0.83277375
P62910	RPL32	1	0.999991453	0.000800151	0.024261972	0.97492933
Q9Y4K3	TRAF6	1	0.999999999	4.75E-23	3.59E-07	0.99999964
P62829	RPL23	1	0.999985204	0.004582717	0.020134687	0.9752678
P62899	RPL31	1	0.99999999	0.000790893	4.64E-06	0.99920446
P31749	AKT1	1	1.000000036	1.05E-25	3.64E-08	1
P04406	GAPDH	1	0.999999951	2.40E-08	0.000953777	0.99904615
P62888	RPL30	1	0.999999915	0.001594319	0.000552396	0.9978532
P08238	HSP90AB1	1	1	1.21E-31	1.91E-10	1
P08670	VIM	1	0.999999966	6.99E-15	2.58E-05	0.99997413
P63104	YWHAZ	1	1	0	7.78E-13	1
P07900	HSP90AA1	1	1	0	7.27E-16	1
P62873	GNB1	1	1	0	1.42E-17	1
P03886	MT-ND1	1	1.000000008	2.32E-19	7.71E-09	1
P03905	MT-ND4	1	1	0	1.42E-24	1
P62258	YWHAE	1	1	0	0	1
P23396	RPS3	1	0.991243486	0.005228526	0.34348196	0.642533
P11142	HSPA8	1	1.000000031	3.29E-16	1.41E-05	0.99998593
P11021	HSPA5	1	0.999999971	2.21E-14	0.000206021	0.99979395
P60709	ACTB	1	1	4.01E-26	1.20E-11	1

P39019	RPS19	1	0.999838986	3.28E-05	0.12615493	0.8736513
P02511	CRYAB	1	0.99910665	0.16285683	0.06690952	0.7693403
P27361	MAPK3	1	0.999999917	7.63E-10	0.009924866	0.99007505
Q92731	ESR2	1	0.999999862	9.21E-10	0.010534761	0.9894651
Q9Y3U8	RPL36	1	0.99781079	0.0065518	0.24068935	0.75056964
P13196	ALAS1	1	0.97037526	0.06133436	0.3179493	0.5910916
P17612	PRKACA	1	0.999999932	2.01E-09	0.00866673	0.9913332
P02792	FTL	1	1	0	1.34E-27	1
P28482	MAPK1	0.999999999	0.999999984	3.34E-11	0.001600544	0.99839944
P17252	PRKCA	0.999999996	0.999999569	7.40E-09	0.014845562	0.985154
P46783	RPS10	0.999999987	0.995875641	0.000533121	0.34329182	0.6520507
P10109	FDX1	0.99999977	0.90154841	0.2342498	0.26630494	0.40099367
Q02156	PRKCE	0.999999645	0.999277559	2.10E-05	0.22561006	0.7736465
P61764	STXBP1	0.999999614	0.97201257	0.07328749	0.30121058	0.5975145
P02794	FTH1	0.999999594	1.000000014	0.005375701	1.39E-08	0.9946243
P06493	CDK1	0.999999009	0.999999969	1.19E-09	0.005060668	0.9949393
P62277	RPS13	0.999998684	0.990720631	0.019664261	0.30100527	0.6700511
P62263	RPS14	0.999998589	0.985545378	0.010100848	0.36621773	0.6092268
P61978	HNRNPK	0.999993521	0.999999581	2.77E-08	0.012584304	0.98741525
P27348	YWHAQ	0.999992721	0.999996872	1.42E-07	0.03330363	0.9666931
P38646	HSPA9	0.999983355	0.998883981	0.000517401	0.21728358	0.781083
P34897	SHMT2	0.999933338	0.99261569	0.00064801	0.37570363	0.61626405
P16220	CREB1	0.999932412	0.999922956	5.93E-06	0.109651975	0.89026505
P31689	DNAJA1	0.999931593	0.91712475	0.2576459	0.24717715	0.4123017
P61088	UBE2N	0.99990959	0.999985302	1.77E-06	0.054590233	0.9453933
P32119	PRDX2	0.999908121	0.90732227	0.15450673	0.31131312	0.44150242
P27469	G0S2	0.99990395	0.999693785	0.001688285	0.1457992	0.8522063
P37840	SNCA	0.999883605	0.998998074	0.001011364	0.21501401	0.7829727
P05388	RPLP0	0.999880692	0.997934177	0.001234627	0.26085025	0.7358493

P61586	RHOA	0.99976098	0.999999894	8.35E-09	0.006734735	0.99326515
P60660	MYL6	0.999697851	0.999949612	0.006963168	0.035640184	0.95734626
O00571	DDX3X	0.999668441	0.97624079	0.05099383	0.31261766	0.6126293
Q96CW1	AP2M1	0.99966822	0.999316248	0.000661518	0.18863103	0.8100237
P35613	BSG	0.999633348	0.92004317	0.08476377	0.3630782	0.4722012
Q04917	YWHAH	0.99963223	0.999999942	7.64E-09	0.005168035	0.9948319
P05387	RPLP2	0.999547902	0.99714983	0.0102467	0.24321783	0.7436853
Q06124	PTPN11	0.999521063	0.999545487	5.44E-05	0.18428978	0.8152013
O95298	NDUFC2	0.998808565	0.81023782	0.26111388	0.2582412	0.29088274
P05089	ARG1	0.997750352	0.9516774	0.2563116	0.22497137	0.47039443
P29353	SHC1	0.995722921	0.99711991	0.00040873	0.30066308	0.6960481
P07437	TUBB	0.994621972	0.999918515	6.96E-06	0.109323055	0.8905885
P40227	CCT6A	0.994405989	0.84761175	0.23256433	0.27412495	0.34092247
Q00610	CLTC	0.993574403	0.997411383	0.003226503	0.23527408	0.7589108
P17302	GJA1	0.991859889	0.987213823	0.007487823	0.3655512	0.6141748
Q71U36	TUBA1A	0.98968188	0.999999947	5.43E-08	0.003102692	0.9968972
O75438	NDUFB1	0.989455788	0.94440108	0.33612338	0.2018933	0.4063844
O75394	MRPL33	0.985481657	0.89929885	0.10834575	0.35172117	0.43923193
P63010	AP2B1	0.976522491	0.80763483	0.2553246	0.25797287	0.29433736
P20336	RAB3A	0.968940329	0.999345566	6.32E-05	0.21154806	0.7877343
P09874	PARP1	0.968900796	0.89460482	0.24892785	0.25588205	0.38979492
Q13617	CUL2	0.954469458	0.991829652	0.002598552	0.3534712	0.6357599
Q9NUB1	ACSS1	0.942897711	0.88279531	0.25409317	0.25925747	0.36944467
Q9Y230	RUVBL2	0.920219249	0.982312106	0.012544926	0.36720082	0.60256636
Q02978	SLC25A11	0.903637399	0.7898282	0.24266653	0.26638642	0.28077525
P14618	PKM	0.901668171	0.999738359	0.000156629	0.13741288	0.86216885
P09912	IFI6	0.87805275	0.93937254	0.10498691	0.32700288	0.50738275
P40305	IFI27	0.862024921	0.977367819	0.018280359	0.37190232	0.58718514
P07203	GPX1	0.653971244	0.90615803	0.12520048	0.33284873	0.44810882

Q9NR31	SAR1A	0.630566157	0.92588316	0.08430742	0.35803857	0.48353717
Q02543	RPL18A	1	0.978703884	5.41E-08	0.63445306	0.34425077
P62241	RPS8	1	0.634337954	1.63E-06	0.5517032	0.08263312
Q3ZCQ8	TIMM50	0.999067678	0.73682338	0.19814903	0.2809093	0.25776505
P30044	PRDX5	0.994719528	0.75822708	0.1388764	0.32947758	0.2898731
Q9HCE7	SMURF1	0.878981221	0.586300558	0.006126368	0.44373107	0.13644312
P27824	CANX	0.839058218	0.612611943	0.022769403	0.4203245	0.16951804
P12236	SLC25A6	0.790992624	0.69991874	0.08530321	0.36749893	0.2471166
O75323	GBAS	0.720103105	0.76596453	0.17177685	0.30557305	0.28861463
P06576	ATP5B	1	0.964643006	0.5645618	0.111708686	0.28837252
P18124	RPL7	1	0.998495747	0.49877954	0.046401337	0.45331487
P49207	RPL34	1	0.995886034	0.6135136	0.056968864	0.32540357
P20674	COX5A	0.999999999	0.970462341	0.90011185	0.020452961	0.04989753
Q9UII2	ATPIF1	0.999999968	0.997853578	0.9844714	0.001937657	0.011444521
O14561	NDUFAB1	0.999999935	0.992212318	0.95897776	0.006914065	0.026320493
Q16718	NDUFA5	0.99999987	0.988687222	0.9562464	0.007690964	0.024749858
P18077	RPL35A	0.999999856	0.999856757	0.96011597	0.000952092	0.038788695
P42765	ACAA2	0.999999771	0.846642464	0.6671988	0.07641258	0.103031084
Q9H3K2	GHITM	0.999999745	0.968166363	0.9254388	0.013395749	0.029331814
P36542	ATP5C1	0.999999698	0.85638204	0.48178276	0.1578698	0.21672948
O95573	ACSL3	0.999999673	0.903963418	0.86692846	0.015700966	0.021333992
P60174	TPI1	0.999999231	0.992031295	0.95360166	0.007948794	0.030480841
O95563	MPC2	0.999998709	0.996872045	0.9898324	0.00125087	0.005788775
O95864	FADS2	0.999998027	0.83357353	0.34142563	0.21895087	0.27319703
P07602	PSAP	0.999995711	0.999999548	0.96548915	1.01E-05	0.034500297
P52758	HRSP12	0.99998862	0.702745026	0.64659274	0.030425994	0.025726292
P51970	NDUFA8	0.999965848	0.986907368	0.9559201	0.007748404	0.023238864
O43674	NDUFB5	0.99996559	0.974666832	0.8625123	0.030141722	0.08201281
Q9P015	MRPL15	0.999960817	0.967593416	0.90844643	0.017981716	0.04116527

O75390	CS	0.999959115	0.96233335	0.8513451	0.03301527	0.07797298
P80404	ABAT	0.999904659	0.818368235	0.70205593	0.052943215	0.06336909
P09669	COX6C	0.999899734	0.995711025	0.9329122	0.010035247	0.052763578
P07919	UQCRH	0.999868951	0.981966924	0.9230066	0.015123201	0.043837123
Q9BYD1	MRPL13	0.999867505	0.959208468	0.9108585	0.016014798	0.03233517
P82933	MRPS9	0.999834966	0.927828898	0.86067957	0.025103291	0.042046037
Q16775	HAGH	0.999806809	0.983939671	0.9668469	0.004822366	0.012270405
Q5T653	MRPL2	0.999775432	0.923669619	0.8560511	0.025693253	0.041925266
O15235	MRPS12	0.999694901	0.858695402	0.7371213	0.051765192	0.06980891
P14324	FDPS	0.999609711	0.995270333	0.96587366	0.005203997	0.024192676
P07814	EPRS	0.999600351	0.883251846	0.5792243	0.117965326	0.18606222
O95168	NDUFB4	0.999593468	0.984854648	0.9063917	0.018709008	0.05975394
Q9BYD3	MRPL4	0.999582522	0.79711976	0.65005565	0.06798048	0.07908363
P10606	COX5B	0.99956144	0.932710732	0.8324618	0.035698842	0.06455009
P56134	ATP5J2	0.999560279	0.996035919	0.9654403	0.005078142	0.025517477
Q9HAV7	GRPEL1	0.999522541	0.921036153	0.8421767	0.029810553	0.0490489
P24539	ATP5F1	0.999495428	0.941299891	0.862942	0.027436037	0.050921854
P00505	GOT2	0.99946463	0.80807475	0.6786997	0.05923132	0.07014373
Q9NX20	MRPL16	0.999382716	0.952983065	0.9053417	0.016434962	0.031206403
Q9NRX2	MRPL17	0.999279001	0.910274884	0.8269117	0.03246227	0.050900914
P62917	RPL8	0.99926424	0.88005756	0.6610446	0.0870298	0.13198316
Q9BPW8	NIPSNAP1	0.999122184	0.87433542	0.3528	0.21506014	0.30647528
P82912	MRPS11	0.999086643	0.933971436	0.88202286	0.019474162	0.032474414
P08708	RPS17	0.99902075	0.95024328	0.4207524	0.16833153	0.36115935
Q9Y6H1	CHCHD2	0.99899516	0.859900984	0.56307185	0.124488324	0.17234081
P26640	VARS	0.998779312	0.825309016	0.6901285	0.0595768	0.075603716
O75964	ATP5L	0.99871666	0.9823885	0.93505913	0.012414165	0.034915205
Q9Y6C9	MTCH2	0.998700816	0.95755601	0.91616285	0.014102535	0.027290625
P50213	IDH3A	0.998596657	0.899973977	0.81945014	0.032318797	0.04820504

Q9UDW1	UQCR10	0.998571119	0.93399469	0.5453238	0.13386643	0.25480446
P82675	MRPS5	0.998560777	0.76979549	0.64540416	0.06005313	0.0643382
O43236	SEPT4	0.99854745	0.7198733	0.41135725	0.15772296	0.15079309
Q9BUE6	ISCA1	0.998507559	0.962548317	0.92117625	0.013599115	0.027772952
P21549	AGXT	0.998444299	0.975917892	0.9364598	0.011454645	0.028003447
O60783	MRPS14	0.998352894	0.929243749	0.85764325	0.026548425	0.045052074
P56385	ATP5I	0.998343172	0.984722809	0.91975844	0.015751299	0.04921307
P15954	COX7C	0.998307528	0.993456985	0.9631344	0.006004954	0.024317631
P30405	PPIF	0.998258416	0.8104547	0.4327974	0.17363392	0.20402338
Q99832	CCT7	0.998145702	0.8515168	0.4990156	0.15007533	0.20242587
P09417	QDPR	0.998009528	0.774426472	0.6922043	0.040586688	0.041635484
P38606	ATP6V1A	0.997452683	0.77667528	0.43734288	0.16184388	0.17748852
Q5TC12	ATPAF1	0.996798764	0.87689302	0.83500385	0.018780904	0.023108266
P82664	MRPS10	0.996022378	0.82976844	0.50710803	0.14370336	0.17895705
Q9P035	HACD3	0.995965535	0.962695977	0.8916399	0.022095652	0.048960425
P11586	MTHFD1	0.995862881	0.75738943	0.6087643	0.07153441	0.07709072
Q96EY1	DNAJA3	0.995776821	0.76504846	0.5555413	0.09937516	0.110132
P46781	RPS9	0.995760936	0.86450712	0.3699404	0.20778747	0.28677925
P17735	TAT	0.995659512	0.942746657	0.9088279	0.012693172	0.021225585
P55060	CSE1L	0.995052726	0.81498944	0.41097626	0.1821791	0.22183408
Q15388	TOMM20	0.993770878	0.938342936	0.8730464	0.023605568	0.041690968
P22061	PCMT1	0.993657723	0.89045703	0.82204086	0.028551988	0.039864182
P46199	MTIF2	0.993420773	0.869763907	0.81473553	0.024208331	0.030820046
P22830	FECH	0.993401795	0.918385133	0.8558514	0.024301708	0.038232025
Q9H3Z4	DNAJC5	0.993333656	0.88014232	0.7831291	0.040349934	0.056663286
Q9NYZ2	SLC25A37	0.993049642	0.938110637	0.856382	0.028860917	0.05286772
Q14318	FKBP8	0.993018472	0.94785632	0.5767666	0.1193501	0.25173962
Q86SX6	GLRX5	0.992663621	0.79970739	0.4507203	0.16285384	0.18613325
P62495	ETF1	0.992101274	0.84436428	0.44967067	0.17059991	0.2240937

Q9NWU5	MRPL22	0.992064214	0.82307032	0.72731173	0.04402834	0.05173025
Q16891	IMMT	0.991991148	0.76766029	0.4347314	0.16067722	0.17225167
Q8NI37	PPTC7	0.991932033	0.923182754	0.83428353	0.03303172	0.055867504
P13073	COX4I1	0.991908532	0.988485269	0.89813656	0.019388873	0.070959836
O95139	NDUFB6	0.991646466	0.971645058	0.903631	0.019709634	0.048304424
Q969G6	RFK	0.991509372	0.984199357	0.9660455	0.005066886	0.013086971
P17568	NDUFB7	0.990695035	0.934160375	0.8670167	0.024616443	0.042527232
Q9Y3D5	MRPS18C	0.990413465	0.906525413	0.8485254	0.023581076	0.034418937
P56378	C14orf2	0.989819475	0.967905521	0.90183496	0.019862851	0.04620771
P36957	DLST	0.989790029	0.935879489	0.8649041	0.025557581	0.045417808
O94925	GLS	0.989656023	0.703414204	0.520057	0.09328788	0.090069324
P80303	NUCB2	0.989422926	0.999652357	0.99866605	7.73E-05	0.000908994
Q9Y3B7	MRPL11	0.988705415	0.900323527	0.82448465	0.030582732	0.045256145
Q01433	AMPD2	0.988155483	0.80516585	0.45194465	0.16207395	0.19114725
O94903	PROSC	0.987922368	0.973557502	0.9425664	0.009496646	0.021494456
P82914	MRPS15	0.987476972	0.955014271	0.9015805	0.018008264	0.035425507
Q9Y5L4	TIMM13	0.985842862	0.7635568	0.36487526	0.19566758	0.20301396
O75947	ATP5H	0.985662228	0.812809875	0.7125632	0.04661164	0.053635035
P39748	FEN1	0.985518073	0.7845	0.5646535	0.10239811	0.11744839
P18859	ATP5J	0.985057807	0.936726589	0.88279355	0.019980095	0.033952944
P00167	CYB5A	0.984428225	0.961573355	0.92502236	0.012217132	0.024333863
P54136	RARS	0.984069018	0.886944122	0.8115372	0.031244354	0.044162568
P62269	RPS18	0.983643362	0.727247215	0.49597022	0.11580888	0.115468115
Q9NZJ6	COQ3	0.983642468	0.992569415	0.9807178	0.002682389	0.009169226
O14548	COX7A2L	0.983245983	0.968323476	0.9219691	0.014399544	0.031954832
Q9UBX3	SLC25A10	0.981930237	0.742195495	0.57222307	0.083516695	0.08645573
Q9Y3D6	FIS1	0.980299712	0.947681133	0.89300835	0.019199455	0.035473328
O76031	CLPX	0.979929009	0.808380102	0.74335307	0.031143742	0.03388329
P09543	CNP	0.97593113	0.66570571	0.3867969	0.1495476	0.12936121

Q96A35	MRPL24	0.97539132	0.972311727	0.9411239	0.009701977	0.02148585
Q96E11	MRRF	0.973763687	0.954691408	0.92501974	0.010605049	0.019066619
Q96B49	TOMM6	0.968544653	0.916165003	0.84520125	0.027580013	0.04338374
P48047	ATP5O	0.967456736	0.83058316	0.59667826	0.10177458	0.13213032
O96000	NDUFB10	0.96721466	0.976830463	0.9358703	0.011804013	0.02915615
Q9Y6N1	COX11	0.961105312	0.817292633	0.7104544	0.049072336	0.057765897
P00558	PGK1	0.960187728	0.7641828	0.62362146	0.06765278	0.07290856
Q9Y512	SAMM50	0.960067459	0.838768315	0.7946087	0.020997675	0.02316194
Q9NX40	OCIAD1	0.959212845	0.916401963	0.8664334	0.020031095	0.029937468
Q9HD33	MRPL47	0.958944111	0.908648137	0.86249065	0.019041313	0.027116174
P36969	GPX4	0.958938342	0.99976113	0.9991285	4.02E-05	0.000592398
Q9Y5P6	GMPPB	0.956877286	0.806883049	0.7065719	0.046884492	0.053426657
O95182	NDUFA7	0.952662504	0.927328847	0.8496327	0.028920297	0.04877585
Q9P0U1	TOMM7	0.952376976	0.996876843	0.95633805	0.006025333	0.03451346
Q96IX5	USMG5	0.949154598	0.860081125	0.8000126	0.02702897	0.033039555
P54687	BCAT1	0.948124687	0.978533863	0.9573005	0.006395102	0.014838262
O75608	LYPLA1	0.946975076	0.771693716	0.6136024	0.07542585	0.082665466
P82673	MRPS35	0.942080958	0.708976343	0.63985676	0.03686995	0.032249633
Q16774	GUK1	0.939580852	0.921969837	0.86092013	0.02352497	0.037524737
Q9NWT8	AURKAIP1	0.937337152	0.944021897	0.88739926	0.02017002	0.036452617
P49406	MRPL19	0.936445338	0.768791716	0.70486456	0.03231939	0.031607766
Q5HYK3	COQ5	0.935708737	0.981425593	0.96112305	0.005890729	0.014411814
Q9Y291	MRPS33	0.92634618	0.911569929	0.88233864	0.012488227	0.016743062
Q14596	NBR1	0.925156663	0.775766578	0.68950796	0.04203652	0.044222098
O95202	LETM1	0.921411677	0.71059363	0.60236657	0.05566208	0.05256498
Q9Y320	TMX2	0.920743203	0.854683912	0.8012104	0.024496702	0.02897681
Q9UJZ1	STOML2	0.915648147	0.874292181	0.77330804	0.042239327	0.058744814
Q96FC7	PHYHIPL	0.911739706	0.7258185	0.58033913	0.07309326	0.07238611
Q9HB07	C12orf10	0.906317824	0.980418693	0.96384895	0.004986767	0.011582976

P82663	MRPS25	0.906295865	0.777767139	0.7180857	0.03010304	0.029578399
P09110	ACAA1	0.897477439	0.702555305	0.5420502	0.08213022	0.078374885
Q9Y3D9	MRPS23	0.89551333	0.835546004	0.7769891	0.027500274	0.03105663
P0C7P0	CISD3	0.894011326	0.880085488	0.7867447	0.038905684	0.054435104
Q14197	ICT1	0.892908863	0.645576434	0.45904768	0.09997071	0.086558044
Q9NTX5	ECHDC1	0.892366386	0.96318731	0.9083242	0.01744389	0.03741922
O43677	NDUFC1	0.891106557	0.935199	0.8235425	0.03900591	0.07265059
Q8WVC6	DCAKD	0.890800888	0.930028214	0.89696395	0.013109689	0.019954575
Q9H2W6	MRPL46	0.890552064	0.921824172	0.8871525	0.014109984	0.020561688
Q9Y4U1	MMACHC	0.887494748	0.973019859	0.9531256	0.006449508	0.013444751
P30041	PRDX6	0.881335543	0.939182111	0.89736813	0.015677566	0.026136415
Q9NWU1	OXSM	0.866230492	0.904589089	0.88403964	0.009175346	0.011374103
P56181	NDUFV3	0.855572576	0.924639874	0.8385541	0.031971317	0.054114457
P49901	SMCP	0.853794429	0.852027298	0.8362065	0.008006006	0.007814792
Q15005	SPCS2	0.849865176	0.901360019	0.82635635	0.030277431	0.044726238
O00411	POLRMT	0.849211067	0.700765137	0.61670005	0.04405421	0.040010877
P22234	PAICS	0.844293164	0.717275604	0.62144244	0.049489677	0.046343487
O43920	NDUFS5	0.84223548	0.718862777	0.66980314	0.026596863	0.022462774
P36551	CPOX	0.839316435	0.699601429	0.5926145	0.055763405	0.051223524
O75414	NME6	0.837644036	0.74021342	0.61349046	0.06324178	0.06348118
P23919	DTYMK	0.836283774	0.760272076	0.67516315	0.04249355	0.042615376
Q9P0M9	MRPL27	0.832595376	0.884776108	0.8192811	0.027794456	0.037700552
O95900	TRUB2	0.829801767	0.547594812	0.47464713	0.043028574	0.029919108
O95258	SLC25A14	0.82366455	0.78210471	0.3653433	0.19942978	0.21733163
Q13405	MRPL49	0.816120205	0.637316244	0.42264682	0.11655322	0.098116204
Q9GZT3	SLIRP	0.815323301	0.922936451	0.8729574	0.019617373	0.030361678
Q7Z7F7	MRPL55	0.814006527	0.752728653	0.6765044	0.038853247	0.037371006
O14957	UQCR11	0.811862697	0.84967788	0.577967	0.11622169	0.15548919
Q8IYB8	SUPV3L1	0.80547563	0.761885599	0.6673476	0.046563562	0.047974437

P62072	TIMM10	0.804332366	0.766549775	0.61861014	0.07138363	0.076556005
Q9UL15	BAG5	0.802877423	0.73450506	0.5897278	0.07200756	0.0727697
Q9Y2S7	POLDIP2	0.798356137	0.822921587	0.7657796	0.027255112	0.029886875
P35270	SPR	0.794333684	0.661695826	0.43487206	0.120775506	0.10604826
Q99766	ATP5S	0.792894344	0.843395552	0.7987084	0.021168942	0.02351821
Q9NW81	ATP5SL	0.791169581	0.735451911	0.6730162	0.03277297	0.029662741
Q8NFF5	FLAD1	0.788951952	0.943214579	0.9040373	0.014398047	0.024779232
Q9H6R3	ACSS3	0.779506068	0.873342138	0.81051046	0.027262898	0.03556878
Q8IWL3	HSCB	0.774124299	0.928007563	0.88498	0.016780153	0.02624741
P14406	COX7A2	0.770559034	0.947933359	0.93226117	0.006218928	0.009453261
Q9NQ50	MRPL40	0.752255527	0.716518898	0.66446537	0.028133448	0.02392008
O14735	CDIPT	0.751504078	0.898553168	0.81300896	0.034312	0.051232208
Q9NX00	TMEM160	0.750906422	0.76432836	0.3508839	0.20253602	0.21090844
Q9BW91	NUDT9	0.740541518	0.745344225	0.709495	0.019329878	0.016519347
Q9NPL8	TIMMDC1	0.730023459	0.767077704	0.69608176	0.035760604	0.03523534
O00408	PDE2A	0.706240744	0.60035161	0.41142341	0.10451077	0.08441743
Q92934	BAD	0.702448465	0.74354683	0.2693425	0.24129803	0.2329063
Q9NQR4	NIT2	0.698045523	0.904232002	0.8610993	0.018081037	0.025051665
P39656	DDOST	0.694534586	0.72359277	0.42194134	0.15307432	0.14857711
Q96AG4	LRRC59	0.692619523	0.75330493	0.5749117	0.086674616	0.091718614
Q96EL2	MRPS24	0.68984259	0.791297245	0.7408782	0.025417935	0.02500111
Q9GZY4	COA1	0.682649195	0.73913152	0.6516452	0.04468327	0.04280305
O95167	NDUFA3	0.67950088	0.804961981	0.69285977	0.052279823	0.059822388
Q9Y5J6	TIMM10B	0.670093512	0.716732913	0.66417277	0.028384028	0.024176115
Q6IPR1	LYRM5	0.662561542	0.94137401	0.9087096	0.01242048	0.02024393
Q9NRP4	SDHAF3	0.654017316	0.844110753	0.77318877	0.03232197	0.038600013
P31939	ATIC	0.648697411	0.569111463	0.48343268	0.049275953	0.03640283
Q9Y5J7	TIMM9	0.620269469	0.683070792	0.5703568	0.059629142	0.05308485
A1L0T0	ILVBL	0.590399161	0.6175807	0.5396109	0.04377454	0.03419526

O95749	GGPS1	0.589032414	0.767482137	0.67438024	0.04589662	0.047205277
Q96ND0	FAM210A	0.588803701	0.723036599	0.6875015	0.019522138	0.016012961
P49448	GLUD2	0.584449152	0.72225469	0.60040563	0.06161824	0.06023082
Q9Y5J9	TIMM8B	0.584215203	0.726049013	0.6324649	0.048198953	0.04538516
O14874	BCKDK	0.583864984	0.72053466	0.39059737	0.1689627	0.16097459
Q13268	DHRS2	0.581219703	0.859773491	0.7958174	0.028562086	0.035394005
Q8WV93	LACE1	0.563172624	0.831644476	0.79162526	0.01936703	0.020652186
Q7KZF4	SND1	0.558688631	0.759443834	0.5490256	0.100469984	0.10994825
Q9H0R6	QRSL1	0.555869486	0.548132892	0.49644646	0.030898862	0.02078757
Q9P0K7	RAI14	0.529432414	0.70516797	0.40773857	0.15349455	0.14393485
O60238	BNIP3L	0.520541841	0.713961984	0.5350126	0.09056191	0.088387474
Q7Z7K0	CMC1	0.51909254	0.772054375	0.73560554	0.019181686	0.017267149
Q8TEL6	TRPC4AP	0.516953404	0.71021734	0.41088727	0.15381874	0.14551133
Q15119	PDK2	0.514144133	0.742510788	0.668313	0.03807633	0.036121458
O00198	HRK	0.500041187	0.680429584	0.52135944	0.08326562	0.075804524
P31930	UQCRC1	0.456511955	0.663313235	0.46684277	0.103812255	0.09265821
Q8WWR8	NEU4	0.435576091	0.587391673	0.49588835	0.05201459	0.039488733